2

SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

GEORGIA INSTITUTE
OF TECHNOLOGY
ATLANTA, GEORGIA 30332

2

9

Final Summary Report,

for the

United States Army Operational Test and Evaluation Agency
5600 Columbia Pike
Falls Church, Virginia 22041

6

Studies: Investigate Slope of the Learning Curve; Explore Tracking Around Target Edges

15 DAAG39-78-C-0047

Conducted by

The School of Industrial and Systems Engineering
Georgia Institute of Technology

10

Leslie G. Callahan, Jr., Principal Investigator

Russell G. Heikes,

Thomas L. Sadosky

Harrison M. Wadsworth

12 98p.

11 July 1979

Table of Contents

page

I. Nature of the Research Program 1

A. Background 1
B. Thesis Activity 1
C. Contract Support for Army Theses 2

II. Development of 1977-1978 OTEA Research Studies 8

A. Background and Overview 8
B. General Approach 9
C. Specific Tasks 10

III. Studies to Investigate the Slope of the Learning Curve in Operational Tests 12

A. Development of the Research Effort 12
B. Results and Conclusions 15
C. Evaluation of Research 17

IV. Studies to Explore Tracking within the Target Edges 21

A. Development of Research Effort 21
B. Problem Statement 21
C. Results and Conclusions 22
D. Areas of Continuing Research 23

Appendix A. Detection of Group Learning Curves from Operational Test Data 24

Appendix B. Testing for Learning with Small Data Sets 40

Appendix C. A Study of Learning in the Operation of a Viscous Damped Traversing Unit 50

Appendix D. A Tracking Performance Study of Large Dimensioned Targets through an Optical Sight 59

Appendix E. A Study of Tracking Speed and Target Size 66

Appendix F. A Study of Aim Point Uncertainty for Different Target Sizes and Shapes 69

Appendix G. Handbook for Early Detection of Learning in Field Testing 74

I. Nature of the Research Program

A. Background

The School of Industrial and Systems Engineering of the Georgia Institute of Technology began to offer Operations Research/Systems Analysis courses at the graduate level in the mid-1950's. A small number of officers and civilians from the Department of Defense who were pursuing graduate degrees in established areas enrolled in these courses. In 1969 the U.S. Army devel)ped a core curriculum for a formal graduate program in OR/SA, and selected Georgia Tech as one of the two civilian institutions for concentrated use in meeting Army graduate educational needs in this area. In 1972 the School was authorized to award a graduate degree in operations research, MSOR. A number of joint reviews have been made in order to improve the Army OR/SA program requirement. The latest was in November 1976. Sixteen Army personnel entered the program in 1969, and by 1973, the program had peaked with 35 students in residence with approximately 20 graduating each year. Since the mid-60's over one hundred officers have received graduate degrees with heavy emphasis on OR/SA methodologies. At present approximately 10 are enrolled in the program.

B. Thesis Activity

At the academic instructional level, methodological course work is closely interrelated with application and research activities. For most Master's degree candidates, the identification and definition of a thesis topic of interest both to the student and to his research advisor requires a disproportionate amount of time when compared with the course requirements or actual thesis research. One of the important objectives to be realized in this program is the development of readily available research topics relevant to Army needs and objectives and potentially interesting to Army personnel, and to competent, involved research advisors. These availabilities are critical if the Army personnel are to complete an acceptable thesis within the time constraint of the program.

During the 1960's and early 1970's a number of informal contacts were made between students, faculty and Army agencies to generate relevant theses research areas and reliable data sources. A host of agency "shopping lists" for proposed theses were made available to Army students. These efforts proved largely unsuccessful, and less than one-tenth of the theses completed by Army officers prior to 1974 were related to Army needs and problems. This situation was summarized in an October 1973 letter from Dr. Wilbur Payne, then Deputy Under Secretary of the Army, to Georgia Tech approving the revised curriculum programs when he stated:

> "I was very interested in the comments you received from the officer students in response to your Proposal Review memorandum. Of particular interest were their remarks concerning the lack of adequate communication between the Army and students, and the resulting scarcity of appropriate military related thesis topics. This has for some time also been a concern of mine. I believe that something can be done to improve this situation, and would be delighted to work with the Institute toward that goal."

C. Contract Support for Army Theses

The first Army sponsored research which supported Army graduate students at Georgia Tech was provided under a contract from the Army Research Office from January 1970 to 31 March 1972. Under the title of "A Research Program in Operations Research and Management Sciences," the scope of work under this contract called for a general research program with emphasis on research, development and engineering administration, and mathematical programming theory and applications. Specific tasks required that Georgia Tech:

1. Construct, and find procedures for the solution of operations research models in areas important to the Army;
2. Identify potential thesis topics and provide experience in model building and analysis to participants in the Army Operations Research Program;
3. Study the application of the models and procedures of military oriented OR models to civilian life.

This contract was funded at a level of $40,000 from the Army Materiel Command, and supported five Army theses as listed below:

"The Development of a Quantitative Model for Resource Allocation within the Exploratory Development Category of the Army Materiel Command," by John M. Grimshaw, Major, Infantry

"A Model of a Manpower Training System with Applications to Basic Combat Training in the United Stated Army," by John E. Miller, Major, Armor

"Maximal Funnel-Node Flows in an Undirected Network," by Duane D. Miller, Lt. Colonel, Infantry

"An Analysis of the System Effectiveness of a Sequential Manpower Training Model," by David S. Grieshop, Major, Artillery

"Maximal Flow with Gains Through a Special Network," by Anthony M. Jezior, Lt. Colonel, Infantry

As their titles reflect, three of these theses were oriented towards theoretical extensions, and only two were directed at the application of theory to solve Army problems. Consequently there was still a need for a better means to bring together students, faculty, and Army agencies.

During the Fall of 1973 and Spring of 1974 a number of conferences and seminars were held between Georgia Tech faculty, students and Army representatives to improve the relevancy of thesis research. In June 1974 the Army Materiel Systems Analysis Agency contracted to support three officers during the year ending in the Fall of 1975. Since then the contract has been renewed annually, and supported research by five additional officers. The AMSAA contracts supported the officer students by providing special office space, leased computer terminals, and other logistic support at Tech, TDY travel funds, and data sources within the sponsoring agency. In addition the contracts also covered approximately 1/4 time salaries, overhead and limited travel for faculty members for efforts beyond what would otherwise be required for their faculty duties. Actual thesis topics were developed between the individual student, the faculty and the sponsor to assure both Army relevance and academic quality and are listed below:

"An Application of Multivariate Statistical Methods in Developing Operational Usage Patterns for U.S. Army Vehicles," by Randall B. Medlock, Captain, Infantry

"An Analysis of Computer Algorithms for Use in Design of Helicopter Control Panel Layouts," by Sam D. Wyman, Captain, Armor

"An Application of Multivariate Statistical Techniques to the Analysis of the Operational Effectives of a Military Force," by James T. Baird, Captain, Infantry

"An Application of Time-Step Simulation to Estimate Air Defense Site Survivability," by James M. Rowan III, Captain, Air Defense

*"A Mathematical Predictive Model of Arm Strength," by Robert S. Lower, Infantry

"Optimum Assignment and Scheduling of Artillery Units to Targets," by Everett D. Lucas, Captain, Artillery

"An Investigation of Aiming Point Strategies for Field Artillery Against Area Targets," by Lawrence Carl Peterson, Major, Artillery

"Error Budgeting Analysis for Tank Gunnery," by James Shiflet, Captain, Armor

Shortly after award of the AMSAA contract in June 1974 negotiations began with the U.S. Army Operational Test and Evaluation Agency to direct the research efforts of Army officer theses research into the general

*Partially supported by Human Engineering Labs thru AMSAA area of Decision/Risk Analysis applied to Operational Tests and Evaluation with initial emphasis on complex command and control systems. Two separate contracts were awarded in the Fall of 1974 in the following subject areas:

1. "Study to Evaluate the Results of Operational Tests and Evaluation of Complex Command and Control Systems" DA39-75-C-0095

2. "Application of Decision/Risk Analysis in Operational Tests and Evaluation" DA39-75-C-0097

Literature search and problem definition in the two areas began in the Summer of 1974 even though the contracts were not awarded until December 1974. They were conducted on a parallel basis with strong interaction between three faculty members and seven graduate students supported under each contract. Frequent seminars and conferences were held throughout the period until individual thesis topics were developed in January 1975. After the Phase I briefing for OTEA at Georgia Tech in February 1975, the individual officers worked independently with their own thesis advisor and committee until graduation in June 1975. A final summary report was made by the faculty at OTEA headquarters in September 1975. This report in both written and oral form discussed the problem, approach, and results of the individual theses and presented results and recommendations in a more general manner than that presented in individual theses which are cited below:

"A Comparison of the Applicability and Effectiveness of ANOVA with MANOVA for Use in the Operational Evaluation of Command and Control Systems," by Thomas N. Burnette, Jr., Captain, Infantry

"An Application of Fault Tree Analysis to Operational Testing," by Gordon Lee Rankin, Captain, Signal Corps

"A Methodology to Establish the Criticality of Attributes in Operational Tests," by Gary S. Williams, Captain, Armor

"An Application of Multivariate Discriminant Analysis and Classification Procedures to Risk Assessment in Operational Testing," by Edward D. Simms, Jr., Captain, Infantry

"An Application of Simulation Networking Techniques in Operational Test Design and Evaluation," by E. L. Brown, Major, Ordnance

"An Application of Bayesian Analysis in Determining Appropriate Sample Sizes for Use in U.S. Army Operational Tests," by Robert L. Cordova, Captain, Ordnance

"Finding a Minimum Risk Path Through a Network Using Resource Allocation Techniques," by Lawrence G. O'Toole, Captain, Armor

At the conclusion of the first year OTEA contract in 1975 it became apparent that it was impossible to clearly delineate work under two separate contracts from the perspective of literature searches, methodological bases and student or faculty efforts. Consequently, a new contract was negotiated for 1975-1976 under the broader scope of "Studies in Support of the Application of Statistical Theory to Design and Evaluation of Operational Tests" with four independently developed tasks. Four theses resulted from this research program, each devoted to a particular subtask. They are entitled:

"An Application of Multiple Response Surface Optimization to the Analysis of Training Effects in Operational Test and Evaluation," by Vernon M. Bettencourt Jr., Captain, Artillery

"A Cost Optimal Approach to Selection of Experimental Designs for Operational Testing Under Conditions of Constrained Sample Size," by Sam W. Russ Jr., Major, Signal Corps

"An Application of Bayesian Statistical Methods in the Determination of Sample Size for Operational Testing in the U.S. Army," by Robert M. Baker, Captain, Infantry

"A Methodology for Determining the Power of MANOVA When the Observations are Serially Correlated," by Norviel R. Eyrich, Captain, Artillery

An additional one year contract was awarded by OTEA on 4 February 1977 which covered theses work for the 1976-1977 academic year with the same title as the previous contract - "Studies in Support of the Application of Statistical Theory to Design and Evaluation of Operational Tests." There were three separate tasks:

1. Study relating to a method to optimize the use of operational test resources and information derived from sequential operational tests.

2. Study relating to a method of optimizing information gained from a small sample.

3. Study relating to methodology for validating the assumptions of multivariate normality in operational test design.

Three theses on one special topics report resulted from the 1976-1977 contract:

"The Use of Operating Characteristic Curves in the Validation of the Assumption of Multivariate Normality and Determination of Sample Size," by Dwight A. Helton, Lieutenant, Signal Corps

"Studies in Support of the Application of Statistical Theory to Design and Evaluation of Operational Tests"

"A Cost Optimal Approach to Selecting a Fractional Factorial Design," by William F. Friese, Captain, Artillery

"A Comparison of Classical and Bayesian Statistical Analysis in Operational Testing," by P. V. Coyle, Captain, Artillery

"A Test for Multivariate Normality in the Army System Acquisition Process," Special Topics Report by Robert S. Young, Captain, Signal Corps

II. Development of 1977-1978 OTEA Research Studies

A. Background and Overview

Research conducted for OTEA in the 1974-1977 period suffered from the lack of specifically desigued study directors on the OTEA staff. As a result there were frequent breakdowns in problem formulation, in process reviews and data collection. As a consequence, per joint agreement between OTEA and Georgia Tech, Mr. Fred McCoy and Mr. Floyd Hill, OTEA, were designated as study directors to assist the COTR in the technical administration of the 1977-1978 contract effort. Informal discussions began on 3 August 1977 when the study directors and COTR visited Georgia Tech for informal discussions with Army graduate students and interested faculty. Literature search was begun in September 1977 at Georgia Tech, and on 9-10 March 1978 the study directors and COTR revisited Georgia Tech for an in-process review and approval of two specific research tasks. After formal award of contract on 21 March 1978, four Army graduate students began field work, data collection and theses research related to these tasks. The four officers received an M.S. in Operations Research on 10 June 1978 and met graduate school requirements with the following theses which have been provided OTEA by separate cover:

"A Study of Learning in the Operations of a Viscous Damped Traversing Unit," by Geoffrey A. Robinson, Captain, Infantry

"A Tracking Performance Study of Large Dimensioned Targets Through an Optical Sight," by Michael L. Morgillo, Captain, Ordnance

"Learning Curves and Their Applicability to Unit Training Levels in Operational Testing," by Jesse L. Brokenburr, Captain, Ordnance

"Testing for Learning with Small Data Sets," by Kenneth A. Yealy, Captain, Infantry

On 11 August 1978 the faculty investigators made an oral presentation for the OTEA staff in Washington and informally reported on the results of the theses research and collateral work including planned preparation of a handbook for field use by OTEA personnel for the on-site detection of learning in operational tests. In November 1979 four refereed papers were presented at the Seventeenth Annual Army Operations Research Symposium at Fort Lee, Virginia.

B. General Approach

The research problem area was approached by first conducting a survey of the relevant technical literature. Both the current open scientific literature and reference material available through DDC and OTEA were evaluated. A series of group and individual meetings between project faculty and the officer-students involved in the program were conducted in addition to the conferences with the OTEA study directors. The purpose of these meetings was to acquaint the officer-students with the general problem area, to discuss previous research effort both in related fields and conducted specifically for the DOD, and to develop specific proposals for current research related to the general project objectives. The officer-student research proposals must have three features:

1. They must be directed towards a problem area of interest to OTEA, as outlined in the project task statement.
2. They must describe a project that constitutes a reasonable contribution to the profession, so that the requirements of a Georgia Tech Master's thesis are satisfied.
3. They must be within the general area of interest of the faculty and other resource personnel currently available.

Subject to these guidelines, the individual research proposals were then developed by the four officer-students involved in the project. They were approved by the project faculty, and by the Associate Director for Graduate Studies of the School of Industrial and Systems Engineering. These officer-student research proposals were also sent to OTEA for evaluation and feedback. Student-officers made field trips to OTEA headquarters, ARI, HEL and the ARTS group at Fort Belvoir for data collections and support. HEL provided equipment and instrumentation support for the tracking task. Finally the OTEA study directors participated in the oral thesis defense in May 1978.

C. Specific Tasks

The first task was directed at developing a methodology for using a thinly-based learning curve slope to assess status of unit training in operational test and evaluation and is reported in Chapter III. The contractual objective of this effort was to develop a simple, straightforward methodology for calculating the significance of the difference between two slopes, one representing the experimental case and one the control case, when one or both curves are based on a relatively few points and there is variance in the point estimates for both curves. The study naturally divided itself into two parallel efforts, the recognition and mathematical description of a representative learning curve (or set of curves) applicable to training levels in operational testing, and the procedure for computing the statistical significance of the difference between slopes for two such curves.

The second task was primarily directed at developing a hitting performance model for manual line-of-sight optical tracking in the range 500 to 3000 meters with a large size target approximately 4 feet in radius. Under terms of the contract, major attention would be directed at testing the hypothesis used in most guide-to-line-of-sight missile system simulations that the tracking error is normally distributed. Study results under this task are reported in Chapter IV. In addition a number of collateral research studies are reported which were not part of the contract specifications, but are important in attempting to model and understand the performance of a gunner in the line-of-sight tracking situation.

III. Studies to Investigate the Slope of the Learning Curve in Operational Tests

A. Development of the Research Effort

The first subtask, which involved the identification of an appropriate mathematical model of learning curves, was data based. A significant portion of the total effort was necessarily devoted to identification of potential sources of appropriate test results. Also required was the development of a systematic procedure to investigate the data bases identified. Capt. Jesse Brokenburr in his thesis, "Learning Curves and Their Applicability to Unit Training Levels in Operational Testing," which addressed subtask 1, had as his objectives:

1. Collect all available data which might illustrate the presence of unit or crew learning.
2. Analyze these data to detect the presence of learning.
3. When learning was found, to determine the best possible statistical model which would describe the learning.

Certain restrictions were placed on his work which should be considered before looking at his results. These were:

1. The data had to come from an operational testing environment.
2. Tests conducted should involve team or crew tasks and performance objectives.
3. Test reports must provide a means of tracking a team from start to finish. That is, the team performance must be measured over time or ordered trials.
4. Test reports should provide some insight into the background information concerning the data which is relevant to the study.
5. Perhaps the most serious restriction was the limitation of time

available for the study. This limitation made it difficult to look for data not readily available at easily accessed army agencies or in the literature.

Capt. Brokenburr spent as much time as possible collecting test results to ascertain their appropriateness for this research. These test results were obtained from OTEA and ARI in Washington and from Ft. Benning. In addition he searched the literature for any adequately described test results that might be useful. The results of this search netted seven data sets which appeared to be useful enough to analyze. These are:

1. Improved Tow Vehicle, from OTEA
2. Dragon, from OTEA
3. REALTRAIN Validation with Combat Units in Europe, from ARI
4. REALTRAIN Validation for Rifle Squad, from ARI
5. Project Stalk, from OTEA
6. Lightweight Company Mortar System, from OTEA
7. Team training, Experiment VIII, conducted by NAVTRADEVCEN, and obtained from the literature searches.

All other data sets found were judged to be unsuitable for analysis for the purpose of this research. This data search suggests that it would be interesting to design some experiments which might further validate findings of this research.

The procedure used in this research to detect the possible presence of learning during the test sequence was of a two-fold nature. The first step consists of plotting the data points and the second is the fitting of a linear equation and testing the slope to see if it is significantly different from zero. The purpose of the graphical procedure is to quickly detect any patterns in the data which might suggest the presence of learning.

In many cases the data sets were abandoned after viewing these plots. In other cases the plots were suggestive of possible models which could be used to fit the data. If the plots suggested the possible presence of learning, the second step was carried out, i.e., the fitting of a linear model to the data. Again, if the computed slope of the fitted model proved to be not significantly different from zero, the data set was abandoned with the conclusion that there was no apparent learning taking place between trials.

When, as a result of the two previous steps, learning was apparent in the data, a series of nonlinear models was fit to the data. The models considered were:

1. $\hat{Y} = at^{-b}$
2. $\hat{Y} = a[\beta + (1 - \beta)t^{-b}]$
3. $\hat{Y} = \alpha[a^{t-1}] + \beta$
4. $\hat{Y} = ae^{bt}$
5. $\hat{Y} = ae^{b/t}$
6. $\hat{Y} = at^{-b} + c$
7. $\hat{Y} = \dfrac{a}{t + b} + c$

Some of the above models can easily be transformed into linear models, however they are still basically nonlinear using the data in its observed form. Model 2 can be transformed into model 6 by letting $c = \alpha\beta$ and $a = \alpha(1-\beta)$. However, model 2 has been used successfully in the literature. It was introduced for this reason. Data which fits one model would fit the other.

Not all seven models could be fit to all the data sets. Some data sets did not have enough trails to fit models containing more than two parameters. The data plots suggested some models over others for some sets. As many models as seemed appropriate, however, were fit to each remaining data set.

After each model was fit to a data set, i.e., estimates of the parameters were obtained, goodness of fit tests were performed to determine if the data fit the model. In addition the regression sum of squares was determined for each model fit to each data set. This enabled the researcher to reject some models as being inappropriate for some data sets. These results also enabled the determination of the best model or models for each data set and the resulting best overall model.

The second subtask required the investigation of the mathematical characteristics of learning curve models, the development of the statistical properties of the estimates of the parameters of the model (and thus the estimates of its slope), and the identification of an appropriate mechanism for verification of the validity of the method.

In Capt. Kenneth Yealy's thesis, "Testing for Learning with Small Data Sets," the task of comparing two sets of data against each other was addressed through the mechanism of comparing each against an absolute standard. In fact, the underlying problem which gave rise to the original objective of this research would not have been solved by meeting the stated objective. This occurs because two curves of the type described may have the same rate of improvement over some particular interval of time, but approach different asymptotes. By comparing each crew to an absolute standard (of zero learning) this problem is overcome.

B. Results and Conclusions

The results of research on the first subtask indicate that four of the seven proposed models might be appropriate for describing the unit learning curve:

1. $\hat{Y} = at^{-b}$
2. $\hat{Y} = \alpha[\beta + (1 - \beta)t^{-b}]$
3. $\hat{Y} = at^{-b} + c$
4. $\hat{Y} = ae^{bt}$

The second and third models each contain three parameters while the first and fourth contain two. Since the second and third did not perform significantly better than the others they were rejected on the basis of parsimony.

Of the remaining models, the first, a power function, has been used by many other researchers to model the learning phenomenon for individuals. The fourth, an exponential model, has not been used as much. On this basis this research concluded that the power function seemed most appropriate.

The reader is referred to Capt. Brokenburr's thesis for complete details of this research. A condensed version of this, as reported in a paper presented at AORS XVI, is attached in Appendix A.

Consistent with the results found in subtask 1, it was assumed that learning can be described by a performance curve of the form $z = 1 - at^{-b} + \varepsilon$ where ε is $NID(0,\sigma_\varepsilon^2)$. Note that this is the performance curve version of the learning curve at $^{-b}$ + c with the constant, c, in the model set equal to 1. The restriction on this constant implies only a scaling of the more general model, and the following methodology is not dependent on this restriction. While Cpt. Brokenburr's conclusions do not support the need to have the constant term in the model it was included for generality; the two parameter model is a special case of that used. All procedures proposed will be applicable to the two parameter case.

Two linear methods and one nonlinear method were developed to test for learning by examining the rate of learning over several trials. The linear procedures are based on testing the average rate of learning that occurs over several trials. Several methods for estimating the average rate of learning and the variance of the observations, σ_ε^2, were investigated. The best method for estimating the average rate of learning, based on the minimum variance of the estimate, was the linear least squares regression, LLSR method, and the best estimator of σ_ε^2, which resulted in the most powerful test, was computed using the first differences of the observations. In the nonlinear method, estimates for σ_ε^2 and the parameters "a" and "b" are obtained and a test on the degree of nonlinearity of the function is conducted using Beale's measure of nonlinearity. If the degree of nonlinearity is small enough then the confidence interval for the slope at any trial can be evaluated by using linear theory approximations. In a comparison of the two procedures, the linear methods were more powerful tests, however, the nonlinear method was able to provide information on the rate of learning at each trial when the nonlinearity conditions were satisfied and significant learning was detected. The more powerful linear test procedure was the LLSR method, which can detect an average rate of learning over 15 trials of 1% at an $\alpha = .05$ level 95% of the time when the standard deviation is $\sigma_\varepsilon \leq .05$. A demonstration of the methodology using data collected during field tests on a viscous damped tracker is presented in detail.

Complete details are presented in Capt. Yealy's thesis and a condensed version of this, as reported in a paper at AORS XVI, is attached in Appendix B.

In addition, a handbook that details the methodology of the best method in a form appropriate for use by army officers in a field testing environment was developed by the principal investigators, and is presented in Appendix G.

C. Evaluation of Research

The restrictions on the data collection effort, in subtask 1, limited the results in that not many different types of training situations were available. The restriction to operational tests meant that conclusions regarding model adequacy are limited, because of severe sample size restrictions. Many of the conclusions are based on three or four tests. With such a small number of tests, many different models will appear to fit the data.

An excellent procedure was developed to accomplish the second objective, the detection of learning in the data. Again the severe sample size restriction influences the results of this phase of the research however.

The third phase of the research, fitting an appropriate model which would describe learning was carried out using well known and available computer procedures. The data limitations were such, however, that no statistical differences could be found among several of the candidate models chosen. The ultimate model was therefore chosen on the basis of parsimony, in that it contains a minimum number of parameters, it did not perform significantly worse than any of the others and performed better than some.

Data were really not available to this research effort to determine which of the four models listed above is most appropriate. Further research should be conducted to determine this, particularly an analysis of the difference between the power function and the exponential model.

As previously discussed, such research .ould most likely involve the conducting of some actual experiments. These could first be conducted using simulation techniques with a computer analysis. Following this, actual trials of the most likely test conditions and models could be run to verify or clarify the results of the simulation study.

The model for the performance curve used in the second subtask is:

$$y = 1 - at^{-b}$$

and is appropriate only if y is viewed as the percent of maximum performance attainable. This requirement is unwieldy in practice, and it turns out that it is not necessary for the procedures developed. A more general form of the model would be

$$y^* = c - at^{-b}$$

This model allows y* to be the value of the measure of effectiveness (not a percent of the maximum attainable). This model is simply a linear transformation of the previous model, and since the procedures developed are invariant to linear transformations of the data, the proposed methodology is appropriate for the second model. This eliminates the need for identifying the maximum attainable performance and manipulating the data before the proposed methodology is applied. A parallel argument can be made for considering the complementary curve as

$$y = at^{-b} + c$$

instead of

$$y = at^{-b}$$

It is felt that the proposed approximate linear method provides excellent protection so long as the variability of fully learned crews on the task of interest is not too great. This is especially true as tne amount of learning increases. (The reader is referred to the results of the simulation studies reported on pp. 82-90 of Yealy's thesis for a complete description of the protection offered.) Even the more complex (computationally) nonlinear technique did not approach the linear approximation's performance. This was entirely unexpected, as it is generally assumed that the more closely the assumed model fits the underlying process the more reliable the inferences (from the assumed model) are. The only explanation we can offer for this not occurring is that the statistical properties of the parameter estimates in the nonlinear model are so bad that they offset the advantage of having the correct model.

Summarizing, a case has been made for the appropriateness of the adequacy of the power function model for describing performance during the learning process. Based on this assumption a methodology has been developed to quantitatively assess whether learning is occurring across a series of trials of a unit or crew, even if the number of trials is quite small. This methodology is simple to execute, requiring only arithmetic operations. While it has not been shown that the proposed procedure is "best," it has been shown, through simulation studies, to perform better than several other reasonable candidate procedures.

IV. Studies to Explore Tracking within the Target Edges

A. Development of Research Effort

Tracking performance studies typically follow one of two basic approaches. The first approach is the development of mathematical models or describing functions that view human tracking performance as part of a closed loop system. The second approach is to describe statistically the tracking error or differences between the true position of the target and the operator's output. It is this second approach that best describes the procedures followed in the tracking studies reported here.

Tracking studies have traditionally focused on well defined or "point" targets. The use of these targets allows the tracking errors to be easily measured and avoids problems associated with the tracker's uncertainty as to the exact aim point. Since there are many variables that influence tracking performance, this and other simplifications from the real world are often necessary in controlled experiments. The main focus of the tracking experiments reported here was to relax the point target assumption and to evaluate the impact on tracking error.

B. Problem Statement

The primary objective of this study was to determine the magnitude and distribution of error when tracking the unmarked center of mass of large dimensioned targets and to compare this error to that found when tracking targets with marked aim points. Since it was necessary to use trained trackers for the primary task, a second study was conducted that focused on the learning of tracking skills. Two additional experiments were conducted to explore specialized topics related to the primary objective. The first was a tracking speed study using large targets and the second was a study of aim point uncertainty for large targets.

C. Results and Conclusions

The above studies are individually reported in Appendices C through F. In summary, the results are as follows:

1. The standard deviation of tracking error was approximately 57 percent larger using targets without marked aim points compared to targets with marked aim points. There was a slight decrease in standard deviation of error as targets became larger; however, this trend was not considered significant from a practical point of view.
2. The distributions of tracking error varied between conditions and subjects, however, the common assumption of normality seems justified. No tendency for bimodal distributions occurred; a concentration on target edges did not occur.
3. Many error measures were evaluated in terms of learning. The standard deviation of tracking error was the best measure in this context.
4. Common learning curve models such as $y = at^{-b}$ are appropriate for the change in standard deviation of tracking error with learning.
5. Target speed is a significant factor in tracking performance but the interaction with target size is not significant.
6. As targets increase in size and complexity there is more uncertainty as to the exact aim point. This effect may diminish or nullify the advantages of optical magnification when tracking large targets without visible aim points.

D. Areas of Continuing Research

In addition to these direct results the research projects have led into additional interesting areas. First, comparing the results of these studies to other tracking studies it was found that experiments using similar equipment and conditions often give different results for the standard deviation of tracking error. A key to this problem may be related to the fact that different studies are often conducted using different ranges. When using a tracking system, such as the one used in these studies, the "gain" or control sensitivity is a function of range. In most cases the error is measured in terms of angular error and control sensitivity is ignored. This complication is compounded when different magnifications are used, and when the target varies from a point to larger targets with less specific aim points.

The tracking literature discusses both magnification and gain effects, but does not offer sufficient data to resolve these problems. Our present intention is to build a laboratory simulator where gain and magnification can be easily varied. The system will be based on a paper tape tracking concept with computer generated tapes. When this system is complete, further studies will be conducted.

A second area of continuing research involves finding the best auto-correlation type model to describe tracking performance. In the research reported here a simple auto-correlation model was used to evaluate learning and to correct error measures for the fact that samples of error are not random but correlated. Improvements in the present model will lead to better analysis of future tracking experiments.

Appendix A

Detection of Group Learning Curves from Operational Test Data

DETECTION OF GROUP LEARNING CURVES FROM OPERATIONAL TEST DATA

CPT Jesse L. Brokenburr, Department of Mathematics, United States Military Academy.
Dr. Leslie G. Callahan, Jr., Dr. Russell G. Heikes, Dr. Thomas L. Sadosky, and Dr. Harrison M. Wadsworth, Georgia Institute of Technology.

INTRODUCTION

The U.S. Army Operational Test and Evaluation Agency (OTEA) is continually required to assess the impact of the training level of a crew or unit engaged in operational tests. This assessment is of particular importance because OTEA has the mission of assisting in the planning, directing, and evaluation of operational testing required during the materiel acquisition process of all major systems and selected non-major systems. Adequate and thorough operational testing is essential in determining an item or system's operational suitability and logistic support requirements. Additionally through these tests a comparison is made between new materiel and existing equipment being operated under the same or similar mission profile.

Essentially, the assessment of crew or unit training levels has traditionally been limited to qualitative techniques such as administering a proposed training program (with the assumption that the completed training equals a given training level) relying on ARMY TRAINING AND EVALUATION PROGRAM (ARTEP) results or using military judgment. Training data is currently overwhelmingly qualitative, whereas quantitative data is much to be preferred in operational test and evaluation.

It is generally agreed that a performance curve describing the progress of training is an asymptotic "learning curve". However, even though it is generally accepted that the individual "learning curve" follows this assumption and appears to be robust, it cannot be assumed that a representative "learning curve" for a crew or unit has these same properties.

Since operational testing usually involves the comparison of baseline systems to newly developed systems, participants are initially determined to be qualified or trained on the baseline system. Prior to the actual conduct of the test, refresher training and/or contractor training is provided on the new system. Through the use of randomization and test design the effect of learning during the test is generally expected to be lessened.

An "after the fact" analysis of data from various operational test reports, primarily from OTEA and data made available through other training and analysis agencies, is conducted. Hence, the objective of this paper is to determine the existence of a representative learning curve (or set of curves) and develop a mathematical description of this curve applicable to training levels in operational testing. The existence of a representative learning curve could be used to develop improved operational tests and evaluation methodology for training effectiveness.

A review of general learning theory results and the application of learning theory concepts to group/team learning indicated that experiments conducted in the Team Training Laboratory demonstrated that basic principles of individual learning could be applied to the team considered as a single entity [3]. The underlying model has three essential features [10]. "First a team is a functioning entity having an output which depends on a defined input from its members. Second, a team itself can be considered as the module of investigation and its responses as amenable to manipulation without necessary reference to the performance of individual team members. Third, team performance can and will vary as a function of the consequences of responses much the same as the performance of an individual learner. In that context the basic principles of individual learning curve robustness will be assumed and analysis of the empirical data will proceed along that line".

METHODOLOGY

Each data set will be analyzed iteratively utilizing the following five step procedure.

1. Determine graphically if learning patterns exist. Sample data is plotted using consecutive trials versus a specified performance measure/measure of effectiveness (MOE) in order to determine if there are patterns which might suggest that learning can be detected.

2. Fit Linear Model.
Linear regression models are used to screen sample data for suitability and further analysis. The linear model

$$Y_i = \beta_0 + \beta_1 t_i + \epsilon_i, \quad i = 1,2,3,\ldots, n$$

where t is the i^{th} consecutive trial, is used to fit empirical data from various test reports. For a given trial t, a corresponding observation Y consists of the value $\beta_0 + \beta_1 t$ plus an amount ϵ, the increment by which any individual Y may fall off the regression line. β_0 and β_1 are the linear parameters in the model and are unknown as well as ϵ, the error or "noise" component which changes for each observation Y. The least-squares method is used to estimate the parameters β_0 and β_1. This method minimizes the sum of squares of deviations from the true line and is written [2].

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 t_i)^2 .$$

Estimates are chosen for β_0 and β_1 which produce the least possible value of S.

The usual basic assumptions for this model were made.

(1) ε_i is a random variable with mean zero and variance σ^2 (unknown), that is, $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$.

(2) ε_i and ε_j are uncorrelated.

The linear model is fit to develop some idea of the behavior of the performance measure over consecutive trials. When estimates of the parameters β_o and β_I are obtained, a screening process is conducted to look at the slope (β_I) of the fitted model. This screening process is used to determine if there is an indication of learning over consecutive trials. We use the value from the t-distribution table (with the appropriate degrees of freedom) to obtain an estimate at a given level. We compare this value with the ratio given by

$$\frac{\hat{\beta}_I - \beta_{IO}}{\sqrt{MS_{E/S_{xx}}}}$$

where MS_E is an estimate of the variance and S_{xx} is the corrected sum of squares of the trials, and β_{10} is the hypothesized value of β_1. If we hypothesize that no learning is occurring β_{10} is set equal to zero. From this we would get some approximate idea of whether or not the slope is negative.

Since the performance measures in the data collected are time components and measurements of error over consecutive trials, a negative slope for the regression line would indicate that learning is occurring over consecutive trials. If no learning is detected the data is not subjected to further analysis.

3. Fit Nonlinear model.

Upon determining the suitability of the data, that is, graphically detecting discernible patterns and rejecting the null hypothesis in step 2, that the slope of the regression line is zero, nonlinear learning models are fit to the data. These include learning models suggested in the literature and/or variations based on the graphical patterns of the raw data (see Table 3-1). The selection of models is restricted to functional relationships between two variables whereby, the performance measure (Y) can be separated from the trials (t) in such a way that $Y = f(t)$. Using this relationship, the performance measure is considered to be the dependent variable and the consecutive trial is the independent variable.

The SPSS (Statistical Package for the Social Sciences) subprogram NONLINEAR [9] is used to apply nonlinear regression analysis to estimate parameters that appear in the learning model in a nonlinear fashion.

Table 3-1. Learning Models

Model	Origin
$\hat{Y} = at^{-b}$	T.P. Wright [11]
$\hat{Y} = a[\beta+(1-\beta)t^{-b}]$	De Jong [1]
$\hat{Y} = \alpha[a^{t-1}] + \beta$	Pegels [8]
$\hat{Y} = ae^{bt}$ *	*models suggested by graphical patterns in the data [6]
$\hat{Y} = ae^{b/t}$ *	
$\hat{Y} = at^{-b} + c$ *	
$\hat{Y} = \frac{a}{t+b} + c$ *	

The SPSS subprogram NONLINEAR utilizes the Least Squares Estimation function to estimate the unknown parameters by minimizing the error sum of squares. For each case, the performance measure (dependent variable) is defined:

$$Y_i = f_i(t,\theta) + \varepsilon_i, \quad i = 1,2,\ldots,n$$

where $f_i(t,\theta)$ stands for the model function chosen, ε_i is the error term, and θ is a vector of parameter estimates. The assumptions made are $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$. The error sum of squares function can be written as

$$S(\theta) = \sum_{j=1}^{n} [Y_j - f_j(t_j,\vartheta)]^2$$

The program minimizes the sum of squares for the model $f_i(t,\theta)$ by choosing suitable values for the unknown parameters (θ) in the model. This in turn will describe as close as possible the behavior of the dependent variable Y. Marquardt's nonlinear minimization technique is used to estimate the unknown parameters.

After the nonlinear model is fit, a direct examination of residuals is conducted and a lack of fit ratio is computed for comparison with other models.

If the original observations of a sample data set do not conform to the model assumptions made, then a log transform of the model may possibly correct the problem. When a direct examination of the residuals for a model indicates that the error component is multiplicative instead of additive, then the log transform of the model should be computed and fitted to the sample data. For example, the model $\hat{Y} = at^{-b}$ has multiplicative error when expressed $\hat{Y} = at^{-b}\varepsilon$ and additive error when expressed as $\hat{Y} = at^{-b} + \varepsilon$. In the former case the log transform can be specified as $\ln \hat{Y} = \ln a - b \ln t + \ln \varepsilon$ but in the latter case the log transform cannot be specified. The multiplicative error is exemplified when variability becomes a function of the magnitude of the responses such as cases where large errors are linked with large responses. When the log transform model is linear it is fit using step 2, when otherwise specified step 3 is used, and then tested for model adequacy. When comparisons are made between the log transform models and nonlinear models in step 5 of the iterative process, the parameter estimates must be converted in order to compare sum of squares.

4. Test for Model Adequacy.

The learning models chosen to fit to sample data are assumed to be tentatively correct. Under certain conditions we can check whether or not the models are correct. This will be accomplished by testing for model adequacy using a "goodness of fit" test and through a direct examination of residuals. The residual of each trial is defined as the amount by which the actual observed value Y_i differs from the fitted value $\hat{Y}_i$ and can be written as $e_i = Y_i - \hat{Y}_i$. If the learning model chosen is not correct, then the residuals contain both random (variance error) and systematic (bias error) components.

Recall that during operational tests, repeat observations are not taken for each crew across trials. However, all crews are observed at each consecutive trial and are assumed to be similar in structure and training level. Therefore, several crew observations at the same trial t_i are considered "repeat" points in the data. These "repeats" are used to obtain an estimate of σ^2 and represent a measure of the random error between crews. As a consequence, we can test for the "goodness of fit" of our model. The hypothesis tested [2] can be stated:

H_O: The model adequately fits the data
H_1: The model does not fit the data

The test involves partitioning the error or residual sum of squares (SS_E) into the following two components:

$$SS_E = SS_{PE} + SS_{LOF}$$

where SS_{PE} is the sum of squares attributable to random error between crews and SS_{LOF} is the sum of squares attributable to the lack of fit of the model. The pure error estimate of σ^2 is found by computing the contribution to the pure error sum of squares from the ith consecutive trial when there are at least two observations such that

$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are n_1 repeat observations at t_1

$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ are n_2 repeat observations at t_2

. . .

$Y_{k_1}, Y_{k_2}, \ldots, Y_{kn_k}$ are n_k repeat observations at t_k.

The total sum of squares for pure error is calculated as follows:

$$SS_{PE} = \sum_{i=1}^{m} \sum_{u=1}^{n_i} (Y_{iu} - \bar{Y})^2$$

where m is the number of distinct levels of t, n_i is the number of observations at trial i, Y_{iu} is a single observation, and $\bar{Y}$ is the sample mean across a particular trial. The total degrees of freedom associated with the total sum of squares pure error is computed as follows:

$$\text{total degrees of freedom} = \sum_{i=1}^{k} (n_i - 1) = \sum_{i=1}^{k} n_i - k = n_e$$

The sum of squares for lack of fit is computed by substitution $SS_{LOF} = SS_E - SS_{PE}$ with $n - 2 - n_e$ degrees of freedom, where n is the total number of observations [2]. The mean square for pure error is

$$MS_{PE} = \frac{SS_{PE}}{n_e} = \frac{\sum_{i=1}^{m} \sum_{u=1}^{n_i} (Y_{iu} - \bar{Y})^2}{\sum_{i=1}^{k} n_i - k}$$

and is an estimate of σ^2.

The pure error sum of squares is introduced into the analysis of variance procedure and the F-ratio is computed. This ratio, $F = \frac{MS_{LOF}}{MS_{PE}}$ is compared with the 100(1-α)% point of an F-distribution with $n - 2 - n_e$ and n_e degrees of freedom if the normality assumption is satisfied. If the ratio is

(1) Significant, this indicates that the model appears to be inadequate. Attempts would be made to discover where and how the adequacy occurs.

(2) Not significant, this indicates that there appears to be no reason to doubt the adequacy of the model and both pure error and lack of fit mean squares can be pooled and used as estimates of σ^2 [2].

The usual tests which are appropriate in the linear model case are in general, not appropriate when the model is nonlinear [2]. As a practical procedure we can compare the unexplained variation with an estimate of $V(Y_u) = \sigma^2$ but cannot use the F-statistic to obtain conclusions at any stated level. In the absence of exact results for the nonlinear models, we can regard this sum of squares as being based on the total degrees of freedom for residuals/error. In the nonlinear case this does not in general, lead to an unbiased estimate of σ^2 as in the linear case, even when the model is correct.

A pure error estimate of σ^2 can be obtained from the repeat observations as discussed earlier. This provides a sum of squares for pure error (SS_{PE}) with n_e degrees of freedom. An approximate idea of possible lack of fit can be obtained by evaluating $SS_E - SS_{PE} = SS_{LOF}$ and constructing a lack of fit (LOF) ratio by comparing mean squares.

$$MS_{LOF} = \frac{SS_{LOF}}{n - n_e} \quad \text{and} \quad MS_{PE} = \frac{SS_{PE}}{n_e}$$

$$\text{Lack of Fit ratio} = \frac{MS_{LOF}}{MS_{PE}}$$

Draper and Smith [2] state that an F-test is not applicable here but that we can use the value from the distribution table (with the appropriate degree of freedom) as a measure of comparison. From this we would get some approximate idea of how well the learning model fits.

Additionally, the statistical inferences on the model are checked through a direct examination of residuals in order to conclude either (1) the assumptions appear to be violated or (2) the assumptions do not appear to be violated. This direct examination will be done by plotting the residuals (a) overall (b) in time sequence, and (c) constructing histograms of the residuals. Model adjustments are made based on this examination of residuals and a careful examination of outliers (unusual points in the data that are far greater than the rest in absolute value, and perhaps lies three or four standard deviations or further from the mean of the residuals). The errors may be linked to equipment failures or errors in recording the observations.

When adjustments are made, the iterative procedure returns to step 3 and the model is refit and tested for adequacy. At this point another learning model or adjusted model is fit to the sample data and checked for model adequacy.

After fitting all selected models for a particular data sample, a comparison of models is conducted in step 5 and a new data set is introduced at step 1.

5. Selection of "Best" Model.

The criterion for evaluating the fitted learning models and selecting the model that provides the "best" fit to the empirical data will be based on the comparison of (1) the lack of fit ratio and (2) the sum of squares for regression (SS_R, the amount of variation in the model explained by regression). This criterion is used because it is a systematic and quantitative basis for selecting the "best" model.

DATA ANALYSIS

Due to the nature of the study, there were limitations placed on the characteristics of the data required. Those limitations are listed below:

1. Data had to come from an operational testing environment.
2. Tests conducted should involve team/crew tasks and performance objectives.

3. Criterion or measures of effectiveness must be applicable to team/crew tasks within the context of group or team definitions as defined by Glaser, Klaus, and Egerman [3].
4. Test reports must provide a means of tracking a team/crew from start to finish. That is, performance measured over time or consecutive trials.
5. When applicable, test reports should provide some insight into the background information concerning the data relevant to this study, such as measurement error and conditions that may have affected the test results ("noise" in the data).

To demonstrate the application of the methodology discussed previously a data sample is presented and analyzed.

Project Stalk

Twenty-five tank crews operating under conditions of competitive stress and rigidly uniform training were timed in their performance at hitting a stationary target which appeared suddenly as a result of the travel of their tank. Eleven different conditions of tank and fire control conditions were run by each of the twenty-five crews participating in the test. Crews were given instructions to obtain a target hit in a minimum time. Crews were timed in their speed at 'ecognizing the target, loading the round, laying the gun, etc., until a hit was obtained. Two types of test courses were used. On the first type, range and characteristics of the target and tank positions were repeatedly observed by the crews. On the second course none of these factors were known by the crews. The experimental design was such that factors related tɔ differences in training, testing conditions, and crew proficiency could be accounted for when comparing the performance of the five tanks. A summarized description is shown below.

1. Performance Measure - Time of detection to hit on target
2. Characteristics
   (a) Twenty-five crews
   (b) Five types of tanks used
   (c) Each crew was trained on a tank immediately prior to firing it.
   (d) Type of activity - Tank gunnery [4,5].

Two aggregate data sets for both the Test Training Course (TTC) and the Test Course (TC) were developed by combining the data for the 16 crews across the four non-transfer targets and the eleven conditions for each target. This provided a method of tracking the crew performances throughout the test according to the Greco-Latin test design used. The TTC data consisted of 678 observations and the TC data consisted of 674 observations over 44 trials. When the linear model was fit to both data sets in step 2 of the screening process, the following results were indicated.

TTC

Source	d.f.	Sum of Squares	Mean Square
Regression	1	87726.475	87726.475
Residuals	676	2827995.42068	4183.425

$$\text{F-ratio} = \frac{87726.475}{4183.425} = 20.97$$

TC

Source	d.f	Sum of Squares	Mean Square
Regression	1	82522.39281	82522.39281
Residuals	672	2440878.25556	3632.259308

$$\text{F-ratio} = \frac{82522.39281}{3632.259308} = 22.719$$

When compared to the F-distribution value for the appropriate degrees of freedom at the 5 percent level, there was evidence to reject that $\beta_1 = 0$. The confidence intervals around β_1 for both data sets did not include zero. Since the estimates of β_1 were both negative, there was an indication that learning was occurring.

Both data sets satisfied the suitability criteria specified in the screening process; therefore, the nonlinear learning models listed in Table 3-1 were fit to the data.

Initially three models were fit.

(1) $\hat{Y} = at^{-b}$

(2) $\hat{Y} = ae^{bt}$

(3) $\hat{Y} = ae^{b/t}$

First analyze the Test Training Course data. Parameter estimates and a residual sum of squares were obtained by using the SPSS Nonlinear Subprogram.

(1) $\hat{Y} = at^{-b}$ where $a = 86.13708$ $b = -.173043$ $SS_E = 2851060.4$

(2) $\hat{Y} = ae^{bt}$ where $a = 77.2504$ $b = -.01792$ $SS_E = 2822300.5$

(3) $\hat{Y} = ae^{b/t}$ where $a = 51.61$ $b = .31028$ $SS_E = 2906957.3$

To obtain an approximate idea of the lack of fit of the models, a pure error estimate of σ^2 was computed as discussed previously by using the 16 crew observations over each trial.

$$SS_{PE} = \sum_{i=1}^{44} \sum_{u=1}^{n_i} (Y_{i_u} - \bar{Y})^2 = 2339080.18552$$

Since $SS_E = SS_{PE} + SS_{LOF}$, the sum of squares for lack of fit was obtained by subtraction. Using the model $\hat{Y} = at^{-b}$,

$$SS_{LOF} = SS_E - SS_{PE} = 20851060.4 - 2339080.18552$$
$$= 511980.214$$

A lack of fit ratio was obtained by comparing the mean squares.

$$MS_{LOF} = \frac{SS_{LOF}}{n-n_e} = \frac{511980.214}{42} = 12190.00512$$

$$MS_{PE} = \frac{SS_{PE}}{n_e} = \frac{2339080.18552}{634} = 3689.4009$$

$$\text{Lack of Fit ratio} = \frac{12190.00512}{3689.4009} = 3.304$$

The lack of fit ratios for (2) and (3) are shown in Table 4-3. To further test the model for adequacy, a direct examination at residuals was conducted. An overall plot of the average residuals across the 44 trials for the 16 crews was constructed. By visual inspection it appeared that the average residuals at trials 1, 4, and 42 were atypical of the others. The majority of the individual residuals appeared to be ±3 standard deviations from the mean of the residuals at those trials. Even though there were one or two residuals which did not exceed the criteria, it was concluded that the removal of the entire set of observations would not adversely affect the analysis. The model $\hat{Y} = at^{-b}$ appears to fit the data and is selected as the "best" model. Even though De Jong's model and $\hat{Y} = at^{-b} + c$ appear to have a somewhat smaller lack of fit ratio with corresponding larger SS regression, the power function ($\hat{Y} = at^{-b}$) is selected due to parsimony. That is, it has fewer parameters and does not appear to be significantly different from the model $\hat{Y} = at^{-b}$ where $a = 104.595$ and $b = -.26492$.

After fitting and selecting the "best" model we must further examine its adequacy. We compute the residuals $e_j = Y_j - \hat{Y}_j$ and then

Table 4-3. Comparative Results for Fitted Models (TTC)

Model	SS_E	SS_{LOF}	SS_R	Lack of Fit Ratio
$\hat{Y} = at^{-b}$	2851060.4	511980.214	1991541.85	3.304
$\hat{Y} = ae^{bt}$	2822300.5	483220.314	2020301.75	3.119
$\hat{Y} = ae^{b/t}$	2906057.3	567877.114	1935644.95	3.665
$\ln\hat{Y} = \ln a - b\ln t$	374.1706	73.8112	12.50802	3.710
$\ln\hat{Y} = \ln a + bt$	369.6397	69.2803	17.03892	3.48186
$\ln\hat{Y} = \ln a + b/t$	384.5419	84.18246	2.13672	4.23078

SS_{PE} = 2339080.1855 (Nonlinear models)

SS_{PE} = 300.3594 (log transform models)

Table 4-4. Comparative Results for Fitted Models (TTC)
(Adjusted Data)

Model	SS_E	SS_{LOF}	SS_R	Lack of Fit Ratio
$\hat{Y} = at^{-b}$	1527619.0	166437.76	1626287.0	1.856
$\hat{Y} = ae^{bt}$	1529402.9	168221.66	1624503.1	1.876
$\hat{Y} = ae^{b/t}$	1545537.8	184356.374	1608368.2	2.06
$\hat{Y} = \frac{a}{t+b} + c$	1534004.0	172822.575	1619902.0	1.927
$\hat{Y} = a[\beta+(1-\beta)t^{-b}]$	1525856.8	164675.375	1628049.2	1.836
$\hat{Y} = at^{-b} + c$	1526337.5	165156.025	1627568.5	1.842
$\hat{Y} = \alpha(a^{t-1}) + \beta$	1597436.3	266255.06	1556469.7	2.635
$\ln\hat{Y} = \ln a - b\ln t$	310.0763	47.767	19.97871	2.765
$\ln\hat{Y} = \ln a + b/t$	308.7863	46.4774	21.269	2.690
$\ln\hat{Y} = \ln a + b/t$	314.2337	51.925	15.82134	3.005
$\ln\hat{Y} = a' + b't$	.32069	.04804	.76351	2.675

SS_{PE} = 1361181.24 (Nonlinear models)

SS_{PE} = 262.30890 (Log transform models)

SS_{PE} = .27265 (other)

NOTE: Atypical points at trials 1, 6, 42 removed.

estimate and examine their autocorrelation function. The sample autocorrelation function of the residuals is denoted by $\{\hat{\rho}_k(e)\}$ [7]. Again, the average residual across each trial is used. Rather than consider the $\hat{\rho}_k(e)$'s individually, we obtained an indication of whether the first 11 residual autocorrelations considered together indicate adequacy of the model. As a general rule k lag coefficients are examined where $k \leq N/4$. This estimate is obtained through an approximate Chi-square test for model adequacy.

$\hat{\rho}_1(e) = .02758$	$\hat{\rho}_6(e) = -.38102$
$\hat{\rho}_2(e) = -.38909$	$\hat{\rho}_7(e) = -.03358$
$\hat{\rho}_3(e) = -.02111$	$\hat{\rho}_8(e) = .37201$
$\hat{\rho}_4(e) = .38570$	$\hat{\rho}_9(e) = -.16558$
$\hat{\rho}_5(e) = -.34704$	$\hat{\rho}_{10}(e) = -.22597$

$$\hat{\rho}_{11} = .02670$$

Approximate Chi-square statistic

$$Q = (N) \sum_{k=1}^{k} \hat{\rho}_k^2(e)$$

$$k = 11 \text{ lags}$$

Test Statistic $Q = 34.57047$

Comparing Q with a 5 percent value chi-square variable w/43 degrees of freedom, we find $\chi^2_{0.05,43} \approx 59.34$. We conclude that there is no strong evidence to reject the model.

The nonlinear models fit to the Test Course data provided the results shown in Table 4-5 and 4-6 for 674 observations over 44 trials.

The parameter estimates for the two test courses are shown below for both the power function and the exponential models.

<u>TTC</u>

$Y = at^{-b}$

$a = 104.595 \qquad b = .26492$

$Y = ae^{bt}$

$a = 74.1207 \qquad b = -.019076$

<u>TC</u>

$Y = at^{-b}$

$a = 76.3596 \qquad b = .180306$

$Y = ae^{bt}$

$a = 67.5696 \qquad b = -.017967$

Table 4-5. Comparative Results for Fitted Models (TC)

Model	SS_E	SS_{LOF}	SS_R	Lack of Fit Ratio
$\hat{Y} = at^{-b}$	2468607.8	493855.022	2704131.2	3.7513
$\hat{Y} = ae^{bt}$	2440991.9	466239.122	2731747.1	3.5415
$\hat{Y} = ae^{b/t}$	2514968.3	540215.522	2657770.7	4.103
$\ln\hat{Y} = \ln a - b \ln t$	389.61005	105.07188	16.90632	5.5391
$\ln\hat{Y} = \ln a + bt$	381.08081	96.54264	25.43555	5.0895
$\ln\hat{Y} = \ln a + b/t$	404.16331	119.625	2.35305	6.3063

SS_{PE} = 1974752.77787 (Nonlinear models)

SS_{PE} = 284.53817 (Log transform models)

Table 4-6. Comparative Results for Fitted Models (TC)
(Adjusted Data)

Model	SS_E	SS_{LOF}	SS_R	Lack of Fit Ratio
$\hat{Y} = at^{-b}$	513771.03	123782.0216	1321367.97	4.141
$\hat{Y} = ae^{bt}$	496887.26	106898.2516	1338251.74	3.576
$\hat{Y} = ae^{b/t}$	551335.23	161346.222	1283803.77	5.398
$\hat{Y} = \frac{a}{t+b} + c$	547924.68	157935.6716	1287214.32	5.284
$\hat{Y} = a[\beta+(1-\beta)t^{-b}]$	506392.74	116943.7316	1328206.26	3.913
$\hat{Y} = \alpha(a^{t-1}) + \beta$	562010.21	172021.2016	1273128.79	5.755

SS_{PE} = 389989.00838

NOTE: Atypical points removed from data.

A comparison indicates that the TTC model parameters are relatively larger than those for the TC. In addition, the learning factor which is represented by the parameter b appears to be larger for the Test Training Course. Curves for fitted models are shown in figures D-3 thru D-6.

CONCLUSIONS

This paper addressed the problem of determining the existence of a representative group/crew learning curve (or set of curves) and the development of a mathematical description of this curve applicable to training levels in operational testing.

A comparison of the fitted models was conducted by comparing the Lack of Fit ratios and the sum of squares for regression computed for each model. This comparison shows that the following models appear to provide an adequate fit to the data analyzed.

(1) $\hat{Y} = at^{-b}$ (The power function)

(2) $\hat{Y} = a[\beta + (1-\beta)t^{-b}]$ (De Jong's model)

(3) $\hat{Y} = at^{-b} + c$

(4) $\hat{Y} = ae^{bt}$

Since the variation of the power function models (2) and (3) did not appear to provide a better fit to the data, model (1) was selected from the standpoint of parsimony or least parameters. In addition, it cannot be stated conclusively that model (1) provides a better fit than model (4). However, based on a survey of the industrial applications of the power function model as reported in the literature, it was concluded that the model $\hat{Y} = at^{-b}$ does adequately fit the empirical data analyzed and can be used as a representative group/crew learning model for this data.

REFERENCES

1. De Jong, J.R., "The Effects of Increasing Skill on Cycle Time and Its Consequences for Time Standards," Ergonomics, Vol. 1, No. 1, 1957.

2. Draper, N., Smith, H., *Applied Regression Analysis*, John Wiley and Sons, Inc., New York, 1966.

3. Glaser, R., Klaus, D.J., and Egerman, K., *Increasing Team Proficiency Through Training, 2. The Acquisition and Extinction of a Team Response*. Pittsburgh: American Institutes of Research, 1962.

4. Hardison, D.C., et al, "A Partial Analysis of Project Stalk Data with Results of Single Tank Versus Single Tank Duels," Technical Note, BRL-980, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, 1955, AD 310-852.

5. Hill, F.I., et al, "An Assembly of Project Stalk Data," Memorandum Report BRL-745, Vol. I, Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland, 1954, AD 23768.

6. Hoerl, A.E. Jr., "Fitting Curves To Data," *Chemical Business Handbook*, 1st Edition, 1954.

7. Montgomery, D.C., Johnson, L.A., *Forecasting and Times Series Analysis*, McGraw-Hill Book Company, 1976.

8. Pegels, C.C., "On Startup or Learning Curves: An Expanded View," *AIIE Transactions*, Vol. I, No. 3.

9. Robinson, B., "SPSS Subprogram Nonlinear-Nonlinear Regression," Manual no. 433, Vogelback Computing Center, Northwestern University, 1977.

10. Wagner, H., et al, *Team Training and Evaluation Strategies: State of-the-Art*, Technical Report HumRRO-TR-77-1, Human Resources Research Organization, Alexandria, Virginia, 1977.

11. Wright, T.P., "Factors Affecting the Cost of Airplanes," *Journal of the Aeronautical Sciences*, Vol. 3, February 1936.

Appendix B

Testing for Learning with Small Data Sets

TESTING FOR LEARNING WITH SMALL DATASETS*

Capt. Kenneth A. Yealy
Dr. Leslie G. Callahan
Dr. Russell G. Heikes
Dr. Thomas L. Sadosky
Dr. Harrison M. Wadsworth

School of Industrial & Systems Engineering
Georgia Institute of Technology

This study was prompted by the desire of the U.S. Army Operational Test and Evaluation Agency (OTEA) to determine if a crew or unit is fully learned in the operation of a system being evaluated. It is important, due to the nature of operational testing to detect this early in the testing program before additional time and money are expended on results that may not be of use in accurately evaluating the system. To obtain timely information the emphasis is placed on development of a methodology that is applicable in a field environment and is appropriate for small data sets.

It is commonly accepted that human performance during the learning of various types of tasks can be described by a monotonic function, referred to as the "learning curve". Limited studies [3] suggest that this type of model is appropriate for groups of humans acting together, as well as individuals. Thus if performance increases as the unit or crew repeats a task, it would seem obvious that learning is occurring. In practice, however, thereis considerable "noise" or random error superimposed on the learning curve. Thus the approach taken is that of a statistical analysis of the series of data points regarding performance on the system. Pertinent assumptions are:

1. That the data available is (or can be made to be) a function only of the experience of the humans using the system.
2. That the monotonic learning curve is appropriate and performance can be modeled across trials as
$$y_i = 1 - at_i^{-b} + \varepsilon_i, \quad i = 1, 2, \ldots N$$
where y_i is the performance measure, t_i is the trial number, a and b are parameters that depend on the nature of the tasks and ε_i is normally distributed random error with mean zero and variance σ_ε^2.
3. The y_i's are assumed to fall between zero and one. This enabled the researchers to draw inferences based on the magnitudes of the parameters.

METHODOLOGY

If a crew is fully trained a plot of performance against trials would be a horizontal line, i.e., it would have a slope of zero. Statistical tests were developed to determine if a data set came from a system where the slope was non-zero.

*This research was sponsored by the U.S. Army OTEA, under Contract Number D AAG39-78-C-0047.

NON-LINEAR ANALYSIS

Since the performance curve function is non-linear in the parameters a and b, directly estimating these parameters by, say, least squares, and using these estimates in the first derivative of the function was somewhat complicated. In addition, the statistical properties of non-linear estimators are not well developed. However, under certain conditions, as discussed by Beale [2,6], it is appropriate to use linear statistical theory results as approximations when analyzing non-linear estimators. An investigation into the performance function, $y = 1 - at^{-b} + \varepsilon$ indicated that those conditions were met for certain sets of values of a, b, N and σ_ε^2. Thus, a procedure to test for learning is:

1. Estimate the parameters in the performance function and the variance of the process using non-linear least squares estimation.

2. Estimate Beale's measure of non-linearity. If it is not too large proceed to 3. If it is too large, stop.

3. Determine confidence limits for each parameter.

4. Determine confidence limits for the slope at any particular point.

5. If the confidence region does not include zero, it can be concluded that learning is taking place.

A computer program to carry out the entire non-linear analysis was developed. Since this procedure is too complicated to be field expedient alternative approximations were pursued.

APPROXIMATE METHODS

Several methods for estimating the "average" rate of learning across trials were considered. These were examined by finding the variance of each as a function of sample size, and the process variance, σ_ε^2. Since an estimator with small variance would be better, the two estimators with the smallest variances were considered further. The best methods of estimation of "average rate of learning" were:

The Average Consecutive Difference (ACD) Method - the average of the differences between consecutive observation is considered. Let

$$\hat{d}_{ACD} = \sum_{i=1}^{N-1} \frac{X_i}{N-1} = \frac{y_N - y_1}{N-1}$$

where

$$X_i = y_{i+1} - y_i,$$

The variance of $\hat{d}_{ACD}$ is

$$V(\hat{d}_{ACD}) = 2\sigma_\varepsilon^2/(N-1)^2$$

The Linear Least Square Regression (LLSR) Method - a straight line is fit through the observations so as to minimize the sum of the squares of the distances from the observations to the fitted line. Details of such procedures are available in most statistical texts (see for example [5]). The slope of this fitted line can be used as an estimate of the rate of learning across trials giving

$$\hat{d}_{LLSR} = \frac{\sum_{i=1}^{N} (t_i-\bar{t})\, y_i}{\sum_{i=1}^{N} (t_i-\bar{t})^2}$$

The variance of this estimate is

$$V(\hat{d}_{LLSR}) = \sigma_\varepsilon^2 / (\sum_{i=1}^{N} (t_i-\bar{t})^2) = \frac{12\,\sigma_\varepsilon^2}{N(N^2+1)}$$

Since $V(\hat{d}_{LLSR}) < V(\hat{d}_{ACD})$ for all values of $N(\geq 2)$ it may seem that the ACD method should be disregarded. However, if the expected value of each of these estimators is examined, it is found that $E(\hat{d}_{ACD}) > E(\hat{d}_{LLSR})$ for all values of a and b, with the magnitude of the difference being a function of a, b and N. Thus the method which results in the largest value of $\hat{d}/\sigma_{\hat{d}}$ is dependent on the true parameter values and the sample size. These, two methods were investigated further in a simulation study which will be discussed later.

VARIANCE ESTIMATION

Since the variance of the slope estimators is a function of the process variance, it is necessary to estimate this quantity. If no learning is occurring an unbiased estimate of σ_ε^2 is

$$(OBS)^2 = \sum_{t=1}^{N} (y_i-\bar{y})^2/N-1$$

However, this estimate will be inflated if there is learning occurring. Two other estimators that are less biased are

$$(SEX)^2 = (N-1) \sum_{i=1}^{N-1} (X_i-\bar{X})^2/2\,N(N-2)$$

and $$(SER)^2 = (N^2+1)\,(N+2)\,MSE/\,(N^3 - 2N^2 + N+1)$$

where $$\bar{X} = \sum_{i=1}^{N-1} X_i/N-1$$

and MSE is the residual mean square after fitting a simple linear regression line to the data.

Examination of the expected values of $(SEX)^2$ and $(SER)^2$ reveals that if there is no learning taking place these two statistics are unbiased estimators of σ_ε^2. The magnitude of the bias was investigated for all three variance estimators for various values of a, b and N. Based on bias alone $(SEX)^2$ was judged somewhat better than $(SER)^2$ and both were considerably better than $(OBS)^2$. However, this does not consider the variability of these estimators. A small simulation study was carried out to estimate the percent of times $(SEX)^2$ and $(SER)^2$ would provide an estimate of σ_ε^2 of given precision. This simulation study suggested that if the process variability was large (say $\sigma_\varepsilon > .06$) and the initial level of performance was high (say $a < 0.5$), then the $(SER)^2$ estimator would be preferred over the $(SEX)^2$ estimator of σ_ε^2.

SIMULATION STUDY

A simulation study was conducted to evaluate the methods discussed above. The simulator generated values of $y_i = 1 - at_i^{-b} + \varepsilon_i$ $i = 1, 2, \ldots N$ where a, b and N were specified and ε_i was a random variable from a normal distribution with mean of zero and variance of σ_ε^2. These data points were used in each of several methods suggested by the previous analysis (described in detail below) and a decision made, at a chosen level of type I error, α, to accept or reject a null hypothesis that no learning was occurring. This was repeated 1000 times for each set of parameters and the percent of times that the null hypothesis was rejected by each method was recorded. The sets of parameter values used were all combinations of $N = 6, 15$; $\sigma_\varepsilon = 0.03, 0.05, 0.07, 0.09$; $a = 0.1, 0.3, 0.5$; and $b = 0, 0.4, 0.8, 1.2$.

The methods considered were:

1. If

   $$t_o = d_{ACD}/\sqrt{V(\hat{d}_{ACD})} > t^* \alpha, N-2, \quad \text{Reject } H_o$$

   where $(SEX)^2$ is used to estimate σ_ε^2.

2. Same as 1, except $(SER)^2$ is used to estimate σ_ε^2.

3. Same as 1, except a heuristic rule based on the magnitude of $(SEX)^2$ and $\hat{a}$ was used to choose the estimator of σ_ε^2.

4. Same as 1, except

   $$t_o = \hat{d}_{LLSR}/\sqrt{V(\hat{d}_{LLSR})}$$

5. Same as 2, except

   $$t_o = \hat{d}_{LLSR}/\sqrt{V(\hat{d}_{LLSR})}$$

*These refer to the (1-α) percentile of the student - t distribution with stated degrees of freedom.

6. Same as 3, except

$$t_o = \hat{d}_{LLSR}/\sqrt{V(\hat{d}_{LLSR})}$$

7. The non-linear method.

The results of this study* indicated that method 4 in the above list was best. That is, the estimate of the average rate of learning should be the slope of the least squares regression line, but rather than use the regression residual mean square as an estimate of variance we should use the variance estimate based on the differenced series. Tables 1 and 2 compare the best approximate procedure and the non linear procedure. These tables can be interpreted as follows, e.g. for $N = 6$, $\sigma_\varepsilon = .03$, $a = 0.3$ and $b = 0.8$, the method described in 4 will detect learning 98.3% of the time while the non-linear method will detect it 39.9% of the time.

A NUMERICAL EXAMPLE

The procedure identified as most powerful was applied to data that was obtained on a study of performance usin; a viscous damped tripod. The experimental results are given in the second column of "able 3. These values are scaled to increasing values between zero and one (see column 3) to be consistent with the coding assumed in the study, however this is not necessary in practice as the procedure developed is invariant to linear transformations.

Following Method (4) yields the following:

H_0: slope ≤ 0

H_1: slope > 0

Compute:

$$t_0 = \frac{\hat{d}_{LLSR} - 0}{\sqrt{\frac{12\hat{\sigma}_\varepsilon^2}{N(N^2+1)}}}$$

If $t_0 \leq t_{.05,4}$ do not reject H_0

If $t_0 > t_{.05,4}$ reject H_0

Compute an estimate of σ_ε^2 using $(\text{SEX})^2$

$$\hat{\sigma}_\varepsilon^2 = \frac{(N-1) \sum_{i=1}^{N} (x_i-\bar{x})^2}{2N\ (N-2)}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{5(.033449)}{2(6)\ (4)} = .0034843$$

Compute an estimate of the slope using the LLSR method:

$$\hat{d}_{LLSR} = \frac{\sum_{i=1}^{6} (t_i-\bar{t})y_i}{\sum_{i=1}^{6} (t_i-\bar{t})^2}$$

$$\hat{d}_{LLSR} = \frac{1.458}{17.5}$$

$$\hat{d}_{LLSR} = .0833$$

Compute the test statistic, t_0:

$$t_0 = \frac{\hat{d}_{LLSR}-0}{\sqrt{\frac{12\hat{\sigma}_\varepsilon{}^2}{N(N^2+1)}}}$$

$$t_0 = \frac{}{\sqrt{\frac{12(.003484}{6(36+1)}}}$$

$$t_0 = 6.07$$

Since $t_0 > t_{05,4}$ we reject H_0 and conclude learning is occurring during these 6 trials.

TABLE 1

Comparison of the Percent of Significant Tests for Learning Using the Nonlinear Procedure, t_{NL}, and the LLSR Procedure, t_R. The results are based on 1000 simulation runs for each combination of a, b, N, and σ_ε. Tests were conducted at $\alpha = .05$ level.

N = 6 $\sigma_\varepsilon = .03$

	b=0	b=.4	b=.8	b=1.2
a=.5	$t_{NL}=.080$ $t_R=.067$	$t_{NL}=1.000$ $t_R=.998$	$t_{NL}=.895$ $t_R=1.000$	$t_{NL}=.623$ $t_R=1.000$
a=.3	$t_{NL}=.071$ $t_R=.059$	$t_{NL}=.938$ $t_R=.913$	$t_{NL}=.399$ $t_R=.983$	$t_{NL}=.200$ $t_R=.982$
a=.1	$t_{NL}=.053$ $t_R=.056$	$t_{NL}=.079$ $t_R=.318$	$t_{NL}=.025$ $t_R=.415$	$t_{NL}=.025$ $t_R=.497$

TABLE 2

Comparison of the Percent of Significant Tests for Learning Using the Nonlinear Procedure, t_{NL}, and the LLSR Procedure, t_R. The results are based on 1000 simulation runs for each combination of a, b, N and σ_ε. Tests were conducted at the $\alpha = .05$ level.

N = 6 $\sigma_\varepsilon = .05$

	b=0	b=.4	b=.8	b=1.2
a=.5	$t_{NL}=.080$ $t_R=.086$	$t_{NL}=.940$ $t_R=.958$	$t_{NL}=.438$ $t_R=.997$	$t_{NL}=.187$ $t_R=.995$
a=.3	$t_{NL}=.072$ $t_R=.080$	$t_{NL}=.439$ $t_R=.715$	$t_{NL}=.103$ $t_R=.912$	$t_{NL}=.056$ $t_R=.894$
a=.1	$t_{NL}=.015$ $t_R=.084$	$t_{NL}=.008$ $t_R=.257$	$t_{NL}=.005$ $t_R=.328$	$t_{NL}=.016$ $t_R=.381$

TABLE 3. EXAMPLE CALCULATIONS

Trial Number	Result	$V = y_i$	X_i	$(X_i - \bar{X})^2$	$(t_i - \bar{t})y_i$
1	4.2124	.4260	.1328	.003564	-1.0650
2	3.4920	.5588	.0778	.000022	- .8382
3	3.0702	.6366	.1582	.007242	- .3183
4	2.2126	.7948	.0740	.00001	.3974
5	1.8113	.8688	-.0773	.022620	1.30332
6	2.2306	.7915			1.9789

CONCLUSION

A computationally simple statistical procedure was developed to test for learning. Simulation results provided evidence that the proposed procedure was preferred to several alternatives considered with respect to ability to detect learning, and that it was proficient at detecting learning in many cases, even with small sample sizes.

*For complete results the reader should consult [10].

REFERENCES

1. Barnes, Ralph M. and Harold T. Amrine, "The Effect of Practice on Various Elements Used in Screwdriver Work," Journal of Applied Psychology, April 1942, pp. 197-209.

2. Beale, E. M. L., Confidence Regions in Non-Linear Estimation," Royal Statistical Society London Journal, Series B 22-23, 1960-1961, pp. 41-87.

3. Brokenburr, Jesse, "Learning Curves and Their Applicability to Unit Training Levels in Operational Testing," Unpublished M.S. Thesis, ISyE, Georgia Institute of Technology, Atlanta, Ga. June 1978.

4. Caplan, Stanley H. and Walton N. Hancock, "The Effect of Simultaneous Motions on Learning," Journal of Methods Time Measurement, September-October 1963, pp. 23-31.

5. Draper, N. R. amd Smith, H., Applied Regression Analysis, John Wiley and Sons, Inc., 1966.

6. Guttman, Irwin and Meeter, Duane A., "On Beale's Measures of Nonlinearity," Technometrics, Vol. 7, No. 4, dated November 1965.

7. Hancock, Walton M. and Prakash Sathi, Learning Curve Research on Manual Operations: Phase II, Industrial Studies, rev. ed., MTM Association for Standards and Research, Research Report 113A, Fair Lawn, N.J., 1969.

8. Hancock, W. M. and Sathe, P., "Learning Curve Research on Manual Operations," Methods Time Measurement Research Studies, Report 113A, dated 1969.

9. Hartley, H.O., "The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares," Technometrics, Vol. 3, No. 2, May 1961, pp. 269-280.

10. Yealy, Kenneth A, "Testing for Learning with Small Data Sets", Unpublished M.S. Thesis, ISyE, Georgia Institute of Technology Atlanta, Ga. June 1978.

Appendix C

A Study of Learning in the Operation of a Viscous Damped Traversing Unit

A STUDY OF LEARNING IN THE OPERATION OF A VISCOUS DAMPED TRAVERSING UNIT

Capt. Geoffrey A. Robinson
Dr. Thomas L. Sadosky
Dr. Harrison M. Wadsworth
Dr. Russell G. Heikes
Dr. Leslie G. Callahan, Jr.

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Problem Statement

The purpose of this study was to investigate the learning of operators who optically track targets using a viscous damped traversing unit and to find which measures of performance most accurately describes learning.

Equipment and Experimental Procedures

The equipment that was used in this experiment was an apparatus developed by the U. S. Army Human Engineering Laboratory (HEL) at Aberdeen Proving Ground, Maryland. This piece of equipment consisted of a movie camera, lens, rifle scope and tripod. These parts were mounted together as one unit, thus enabling experimenters to make a photographic record to be used in analyzing an operator's ability to track.

The rifle scope was mounted to the top of the movie camera by means of a slide bracket. The scope had a sight extension and a collapsable rubber cuff on the rear to enable the operator to get a good sight picture. The scope had a cross hair for the operator to track the moving target.

The movie camera was a 16mm Milliken camera equipped with a six-inch lens and filmed the moving target at four frames per second. The tripod with its traversing unit weighed approximately 12 pounds. It was designed to be used with loads in the range of 5 to 32 pounds. Such loads typically may be lightweight missile launchers or a variety of optical devices. The eye height relative to ground level may be 22 to 26 inches, depending upon the device affixed to the tripod. In this experiment eye height was adjusted to the individual's position.

The target used was a one meter diameter black circle mounted on a five foot by eight foot white target board. The exact center of the black circle was marked with a white cross. The target board was mounted on an automobile which followed a circular path at a range of 200 meters from the tracking station. The rifle scope was set at five power which presented a visual angle of 85.95 minutes of arc for the black circle. The target moved at a constant speed of five miles per hour or 11 milliradians per second. The experiment was conducted in an open area and during daylight.

Four naive subjects tracked the target for 60 trials where each trial consisted of a minimum of 30 seconds of continuous tracking. The operator rested between trials so as not to interject fatigue into the experiment.

The 60 trials per subject were filmed in the following manner. The first 10 trials of each subject were filmed. Between trials 11 and 20, every other trial was filmed. From 21 through 40, every fifth trial was filmed and the 50th and 60th trials were filmed to complete the data collection.

Data Analysis and Results

The data from the movie film were analyzed on a frame by frame basis using a special projector system that allowed measurement of the horizontal and vertical tracking errors and recorded them on punched tape. Since the task was primarily a horizontal tracking task with little target deviation in the vertical direction, only errors in the horizontal direction are discussed in this paper. Computer statistical programs and plotting procedures allowed various error measures to be calculated. An average was calculated across all four subjects to give a single measure to be plotted vs. successive trial number in order to examine learning. The following measures were used:

1. Reversals - A reversal occurs when the tracker changes from an increasing error status to a decreasing error status or vice versa. The learning curve for reversals (i.e. a plot of the number of reversals per trial vs. trial number) is shown in Figure 1.

2. Crossovers - This represents the number of times the tracker changes from leading to lagging behind the center of the target or vice versa. The learning curve for crossovers is shown in Figure 2.

3. Mean error - This is the average tracking error for the trial. The learning curve is shown in Figure 3.

4. Range of error - This is the maximum range of errors from the target center measured in a trial. The learning curve is shown in Figure 4.

5. Standard deviation of error - This is the standard deviation of tracking errors measured in a trial. The learning curve is shown in Figure 5.

Although learning is evident in several measures the most apparent displays of learning are found in measures of tracking variability rather than average tracking error. The standard deviation of error was selected for further analysis.

Learning Curve Models

Using standard deviation of tracking error three learning curve models with parameter estimates and lack of fit ratios are shown in Table 1.

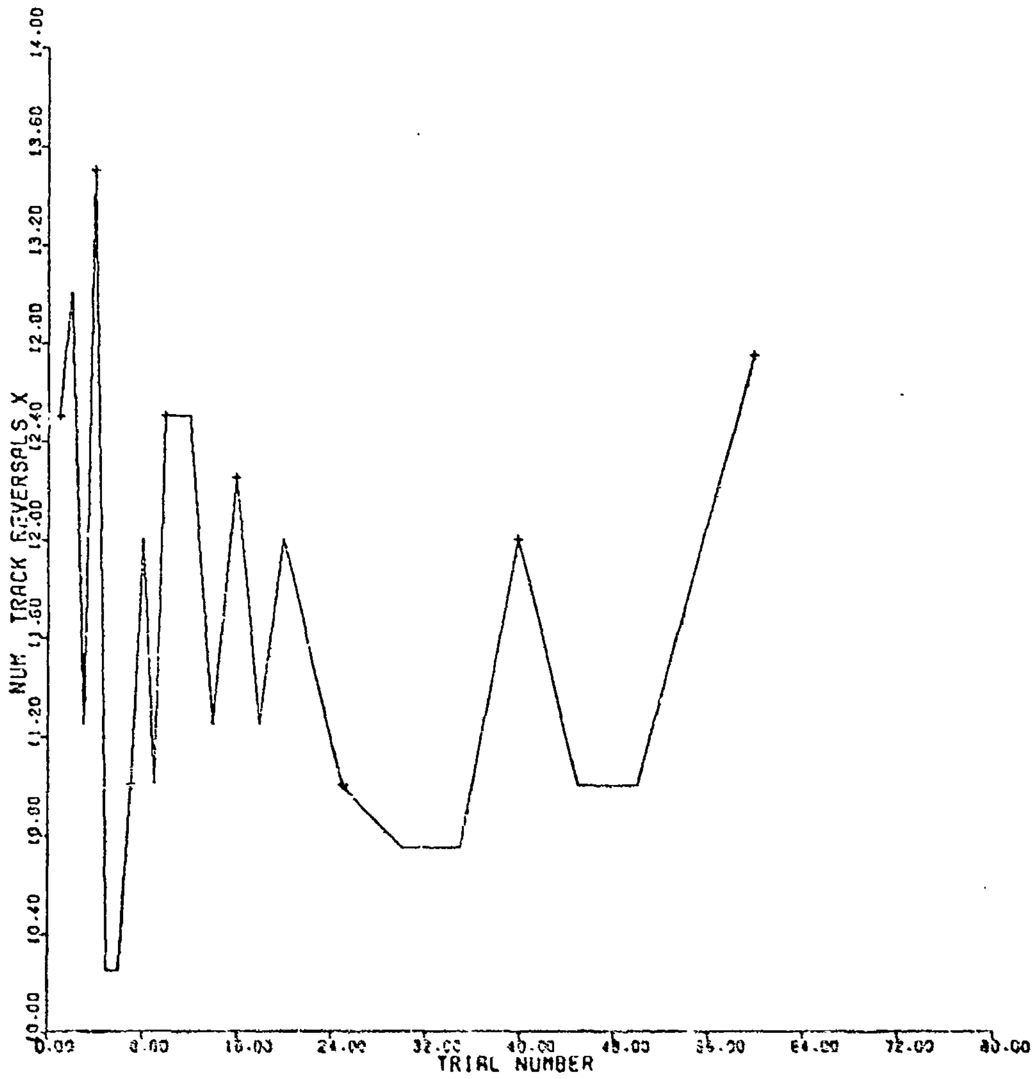


Figure 1

Number of Reversals per Trial vs. Trial Number

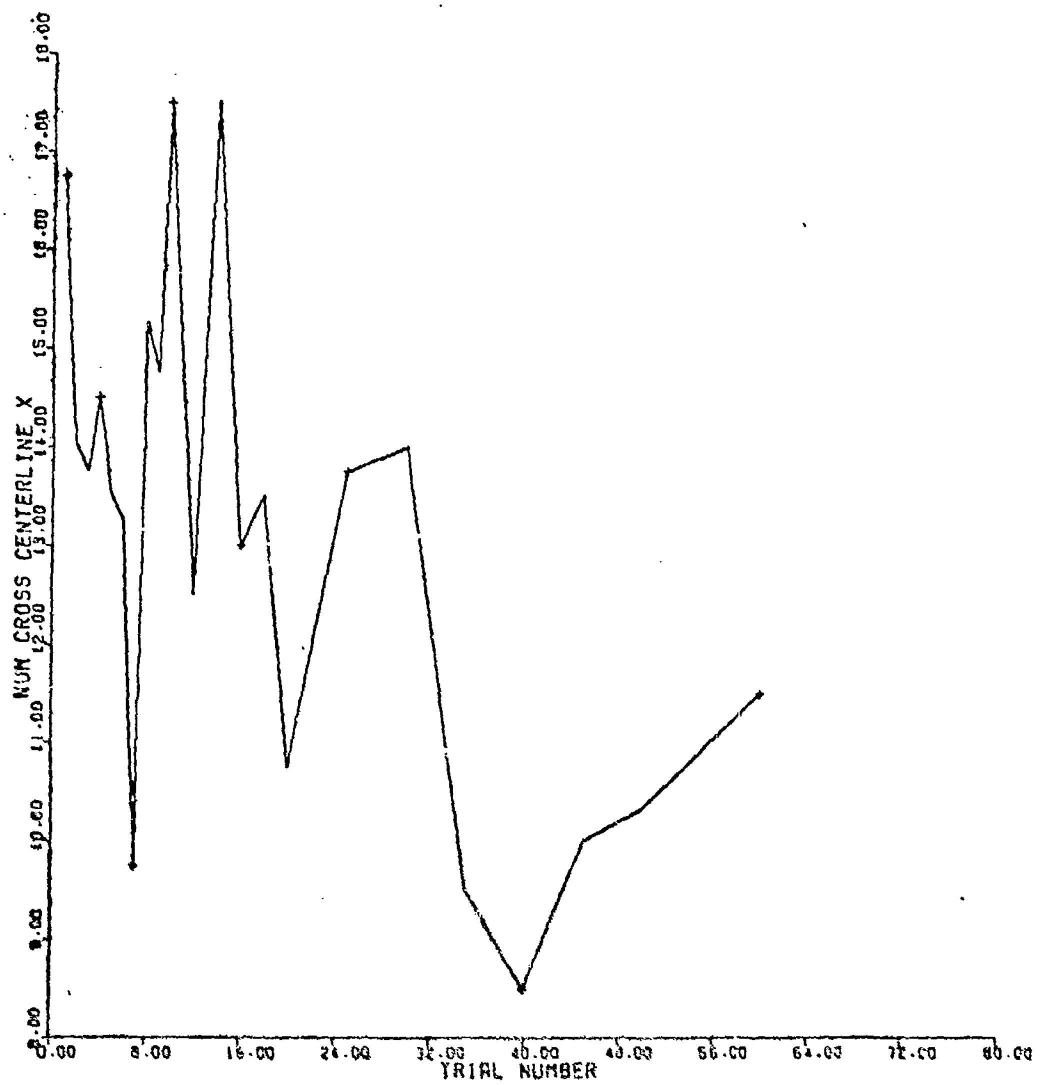


Figure 2

Number of Crossovers per Trial vs. Trial Number

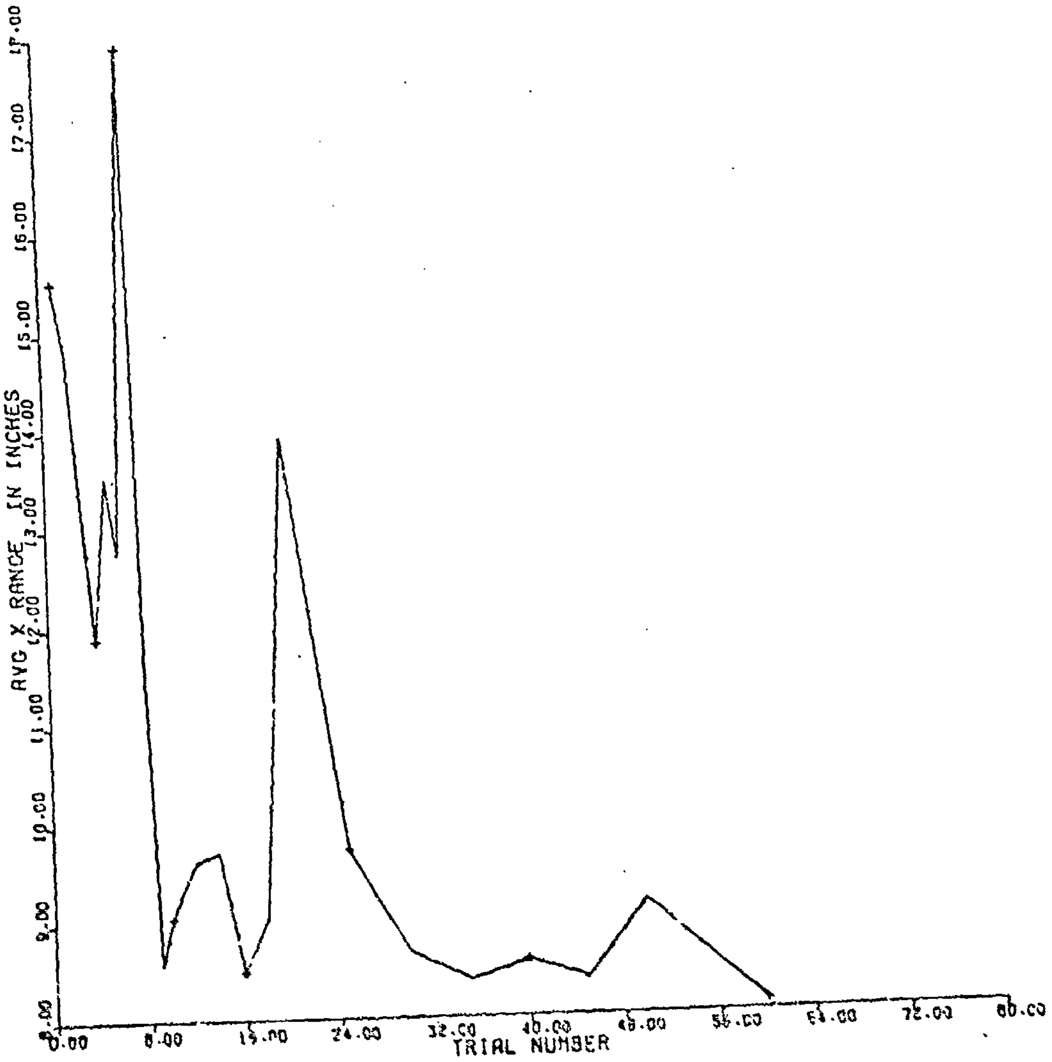


Figure 4

Error Range per Trial vs. Trial Number

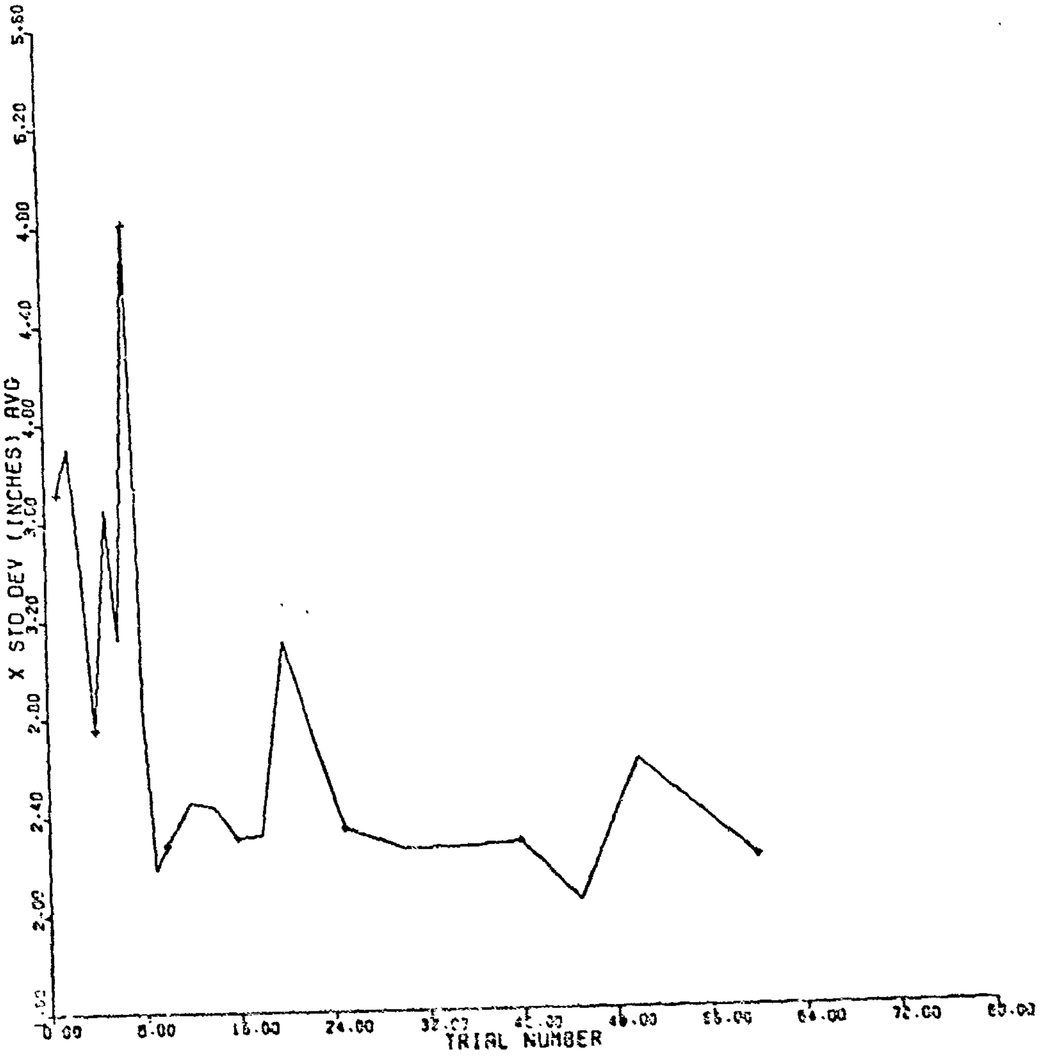


Figure 5

Standard Deviation per Trial vs. Trial Number

Table 1

Learning Curve Models

Model	Sum of Squares of Residuals	Parameter Values		Lack of Fit Ratio
$y = at^{-b}$	171.2	a = 4.01	b = 0.15	.474
$y = ae^{b/t}$	178.6	a = 2.57	b = 0.49	.630
$y = ae^{bt}$	176.3	a = 3.33	b = 0.01	.581

Conclusions

Based on plots of typical performance measures used in tracking, variation of error as measured by the range and standard deviation are better measures of learning than average error, reversals, or crossovers. A large portion of the learning appears to be occuring in the first ten trials.

In fitting typical learning curves to the data of standard deviation vs. trial number, the best of the models tested was $y = at^{-b}$ although the difference between models was not large.

Appendix D

A Tracking Performance Study of Large Dimensioned Targets through an Optical Sight

A TRACKING PERFORMANCE STUDY OF LARGE DIMENSIONED TARGETS THROUGH AN OPTICAL SIGHT

Capt. Michael L. Morgillo
Dr. Thomas L. Sadosky
Dr. Leslie G. Callahan, Jr.
Dr. Russell G. Heikes
Dr. Harrison M. Wadsworth

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Problem Definition

The purpose of the study was to determine the magnitude and distribution of error when tracking the unmarked center of mass of a large diameter circular target and eventually to compare these errors to those found in the tracking of a circular target with a marked aim point at the center of mass.

Scope

The approach to the problem was a search of existing literature to ascertain what work had been accomplished on the subject. Once this had been done, a field experiment, using six subjects, was conducted. A circular target was used and target visual angles between 20 and 200 minutes of arc were investigated.

Equipment

The equipment used in this study was developed by the U. S. Army Human Engineering Laboratory (HEL) at Aberdeen Proving Ground, Maryland. It consisted of a variable power rifle scope (2.5x to 8x) with an extended eye piece. It was affixed by way of a slide mount to a 16 mm. Milligan movie camera. The camera was equipped with a six inch lens and was set to film at a rate of four frames per second.

The camera was secured to a limited production HEL general purpose viscous-damped tripod. The experimental tripod with its traversing unit weighed approximately 12 pounds. It was designed to be used with loads in the range of five to 32 pounds. (A typical military load for this tripod may be a lightweight missile launcher.) The eye height relative to ground level was adjustable from 22 to 26 inches. In this experiment it was set at 22 inches. The traversing unit encompassed a twofold damping system. In the elevation axis, the damping system had a vane type rotor. In the azimuth axis, the system was drum type.

Test Design

The test was designed to encompass visual angles ranging from 20 to 200 minutes of arc, to simulate a tank-size target from ranges of approximately 100 to 3000 meters.

The test condition consisted of a target propelled in the horizontal plane at a constant velocity. Target sizes of one meter and one half meter diameters were used. Two ranges were also used - 100 meters and 200 meters. By varying the power of the scope in conjunction with the two target sizes and two ranges, the desired target visual angles could be achieved (Table 1). The targets were flat black in color and were mounted on a 5' x 8' white target board which was mounted to a vehicle with mounting brackets and tie downs. In an attempt to keep the distance to the tracking station as constant as possible, the target was moved along a relatively flat horizontal, arc shaped path.

After sixty preliminary runs, the subjects were considered trained. Each subject was required to assume a sitting position at the tracking station. A set of pre-printed instructions was read to each subject before the initiation of the experiment. This was done to ensure that all subjects were given identical instructions. Before each individual trial, the subjects were told to lay the rifle cross-hairs on the marked center of the target. A few seconds of film were shot, the mark was removed, and the experimental run was begun. This stationary tracking provided a zero reference point for data reduction and served to eliminate parallax error between the scope and the camera. Additionally, it later served as a medium for determination of experimental human error in data reduction.

Activation of the camera was controlled not by the subject, but by the experimentor who was stationed with the subject at the tracking station. By this method, the subject was not required to concern himself with anything beyond the tracking task.

After initiation of target movement, the target maintained a constant velocity for approximately 45 seconds. To ensure the consistency of velocity time stakes were positioned along the route and the vehicle driver maintained a stop watch count in order to pass the stakes at predetermined intervals. The velocity at 200 meters was five miles per hour, and at 100 meters was 2-1/2 miles per hour. The first five seconds of tracking were devoted to acceleration and initial displacement of the camera, and were not analyzed. Once the tracking began, the subject attempted to track what he perceived to be the center of mass of the target. The test design was blocked to avoid any possible response patterns and balance any additional learning effects.

It should be noted that the experiment was performed outdoors at an unprotected location. The tracker was therefore subjected to the same environmental conditions, such as wind, which would be encountered during the firing of a light weapons system. Experimentation was terminated, however, when strong wind gusts or rain developed.

Conclusions

The conclusions, based on the experimental data, indicated that the shape of the distribution of error did change slightly as a function of target visual angle. In the horizontal plane, the tendency toward a uniform distribution shifted when target visual angle was increased toward a more peaked unimodal distribution. No bimodal distributions or indications of tracking the target edges were found. Practically none of the subject distributions exactly resembled the normal, it is conceivable, however, that a near normal situation could occur if a considerably increased number of data points per run were collected.

It has been shown in the literature that the combination of tracking distributions which are not in themselves normal, often yield a combined resultant distribution which is normal. The frequency histograms derived in this research were not combined by any statistical process, thereby preserving the individual subjects error distributions. It was felt that an examination of these distributions would give a more meaningful comparison of tracking performance on large targets.

In evaluating these error distributions, the following results were obtained. First, the standard deviation of error indicated a decreasing trend from 21.38 to 137.52 minutes of arc; (Figure A) at this point a large increase occurred. Here it should be noted that an actual change in target distance took place. A direct comparison of the results from two ranges should not be made. A linear regression analysis of the first nine points showed a significant, but slight, negative slope.

The time series autocorrelation model was the final attempt at analysis. The results obtained from this model yielded a slight but statistically significant decrease in standard deviation corrected for autocorrelation as visual angle increased (Figure B). This behavior was consistent with the tendancy for the error distribution to become more peaked as the visual angle increased. This can be interpreted as a tendency for the tracker to make fewer corrective motions as target size increases.

In the vertical plane, the expected results were achieved. Since the course was fairly flat, little correction was made in this plane. Throughout all the frequency histograms, a large concentration of points remained around the perceived target center. This remained constant among the range of visual angles and was verified by the lack of significance, at 5 percent, of the regression lines fitted through the plots for the standard deviation of error, standard deviation corrected for autocorrelation, the range and autocorrelation coefficients.

It has been demonstrated that although the trend is statistically significant, the decrease in standard deviation as a function of visual angle is slight. In general, for practical purposes, it appears that the subjects were able to track center of mass of the circular target with very nearly the same "radial error" no matter what the apparent target size. In a concurrent study, using the same conditions and subjects, a trained subject tracked the same targets, but with a marked center point with a standard deviation of error about the point of .2667 milliradians. In this study, the standard deviation of error about the smallest target visual angle was .4195 milliradians.

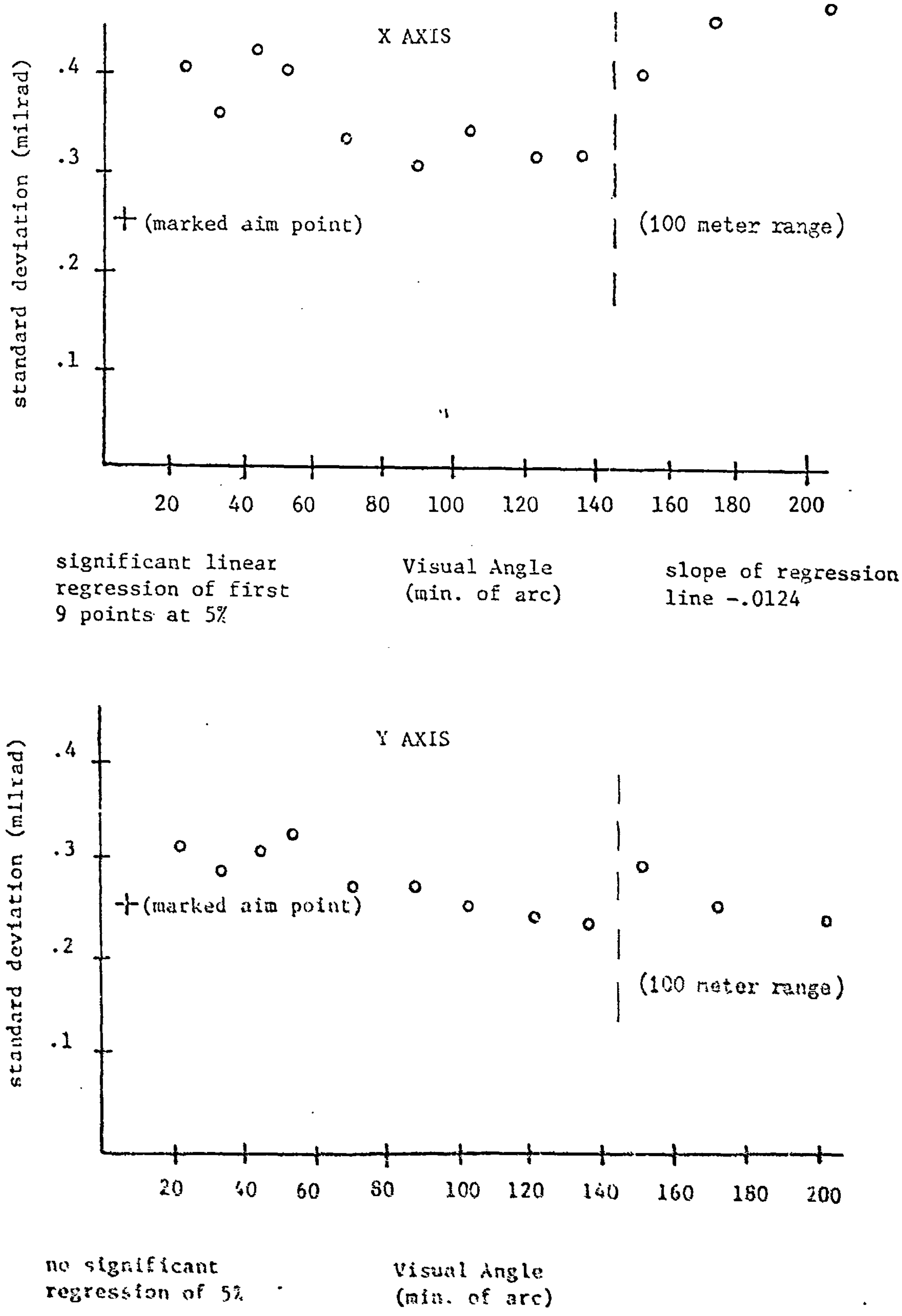
This indicates a substantial, 57 percent, increase in standard deviation of error when a marked aim point is not used. For practical purposes this increase is approximately constant for target sizes ranging from 20 to 200 minutes of arc. The same type of increases are present using the sample range and the standard deviation adjusted for autocorrelation. There was not a significant difference in mean tracking error of targets with marked and unmarked aim points.

Table 1. Experimental Conditions

Range (meters)	Target Size (m)	Scope Power	Visual Angle (min of arc)	Condition
200	½	25x	21.48	1
200	½	4x	34.38	2
200	1	2.5x	42.97	3
200	1	3x	51.57	4
200	1	4x	68.76	5
200	1	5x	85.95	6
200	1	6x	103.14	7
200	1	7x	120.33	8
200	1	8x	137.52	9
100	1	4.5x	154.71	10
100	1	5x	171.90	11
100	1	6x	206.28	12

Velocity was 5 mph. or 11 milliradians per second at 200 meters and 2.5 mph. at 100 meters.

$$\text{Visual Angle min. of arc} = \frac{(53.7)(60)L}{D}$$

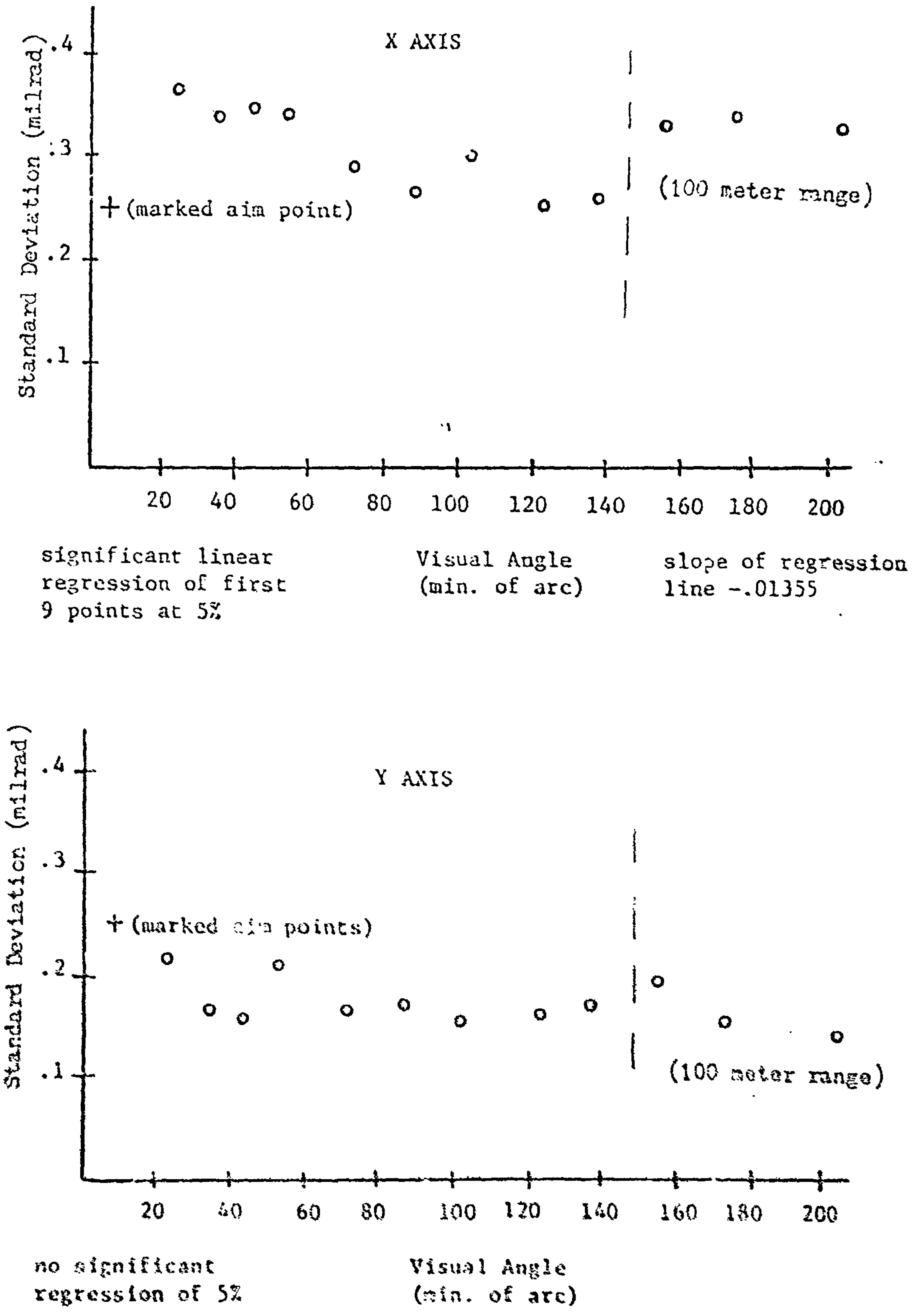


significant linear regression of first 9 points at 5%

slope of regression line -.0124

no significant regression of 5%

Figure A. Standard Deviation of Error



significant linear regression of first 9 points at 5%

slope of regression line -.01355

no significant regression of 5%

Figure B. Standard Deviation of Error Correcte3 for Autocorrelation

Appendix E

A Study of Tracking Speed and Target Size

A STUDY OF TRACKING SPEED AND TARGET SIZE

Dr. Thomas L. Sadosky
Capt. Geoffrey A. Robinson
Capt. Michael L. Morgillo

Problem Statement

It has been well established that tracking speed will have an influence on tracking error. The purpose of this study was to establish the interaction effects of target speed and target size. If this interaction can be shown to be small, the general results found in other target size studies can be extended over a range of tracking speeds.

Equipment and Experimental Procedure

The equipment that was used in this experiment was an apparatus developed by the U.S. Army Human Engineering Laboratory (HEL) at Aberdeen Proving Ground, Maryland. This piece of equipment consisted of a movie camera, lens, rifle scope and tripod. These parts were mounted together as one unit, thus enabling experimenters to make a photographic record to be used in analyzing an operator's ability to track. (For a detailed description of the equipment see appendix C.)

Two trained subjects tracked three different target sizes at four different speeds. The Range was held constant at 200 meters. The targets were solid black circles on a white background and subtended 32, 86 and 138 minutes of arc. The tracking speeds were 2.5, 5, 7.5, and 10 miles per hour. The experiment was a full factorial with 2 replications.

Data Analysis and Results

The data from the movie film were analyzed on a frame by frame basis using a special projector system that allowed measurement of the horizontal and vertical tracking errors and recorded them on punched tape. Since the task was primarily a horizontal tracking task with little target deviation in the vertical direction, only errors in the horizontal direction were analyzed. The standard deviation of tracking error for each of the 48 trials was calculated and an analysis of variance performed on this measure of tracking error. The ANOVA table is given in Table 1.

The results show no significant interaction effects at the .05 level of significance. As expected tracking speed was significant. Target size was not a significant factor which confirms the results of the previous studies on target size. Subject effects were not significant in this study.

The lack of a significant interaction between target size and tracking speed confirms that the general results found in the previous target size studies conducted at only one speed may be generalized over a broader speed range.

SOURCE	DF	MS	F	SIG @ .05
SUBJECT (A)	1	.877	1.87	
TARGET SIZE (B)	2	.548	1.17	
SPEED (C)	3	1.253	2.67	*
A × B	2	.523	1.11	
A × C	3	.222	.47	
B × C	6	.387	.82	
A × B × C	6	.244	.52	
ERROR	24	.468		
TOTAL	47			

Table 1

ANOVA Table for the Tracking Speed Study

Appendix F

A Study of Aim Point Uncertainty for Different Target Sizes and Shapes

A STUDY OF AIM POINT UNCERTAINTY FOR DIFFERENT TARGET SIZES AND SHAPES*

Dr. Thomas L. Sadosky
Mr. Jeff Grant
Mr. Bill Cole

Problem Statement

The purpose of this study was to determine the ability of subjects to locate unmarked aim points on objects of different size and shapes.

Experimental P:ocedure

Four different targets were used in this study: a line, a rectangle, a circle and a silhouette tank. These targets are shown in Figure 1. As shown in the figure, each target was presented in four nominal sizes, a 1, 2, 3, and 4 inch size. This nominal size corresponded to the line length or longest side length for the line, rectangle, and tank. The nominal size corresponds to the diameter for the circle.

The target pages (see Figure 1) were presented to four groups of subjects with four subjects per group. The subjects were instructed to mark the "center" of the target; in the case of the tank a figure showing the desired aim point was shown to the subjects prior to the experiment, but was not available to them during the experiment. Each subject group worked with the targets in a different order, and proceeded through the size ranges in a different order. Two replications of the experiment were conducted generating a total of 512 data points.

Data Analysis and Results

An overlay was used to measure the deviation in inches from the true aim point for all targets. The mean errors for all groups are shown in Table 1. An analysis of learning transfer according to the order in which the subjects marked each type of target showed transfer was negligible.

Table 1 shows that absolute error increased from the line to rectangle to circle to tank. It also shows that working with the tank was a considerable more difficult task - the errors are about twice the magnitude of those on other targets. It is also found that the absolute error increases according to the size oi each target (with the exception of a small deviation between 1 inch and two inches for the circle). The error in terms of percent of nominal size remains fairly constant within one target type.

Conclusions

Two main conclusions are derived from this experiment.

*This study was conducted as a class project in ISyE 6221, Manual Control, at Georgia Tech.

1. Aim point uncertainty, as measured by absolute location errors on fixed targets, increases with target size. In terms of percent of nominal size the errors remain fairly constant for a given target type.

2. Different target types present different difficulty in locating aim points. Irregular shaped objects where the aim point has no specific references or obvious geometric symmetry to aid in its location present the most difficult case.

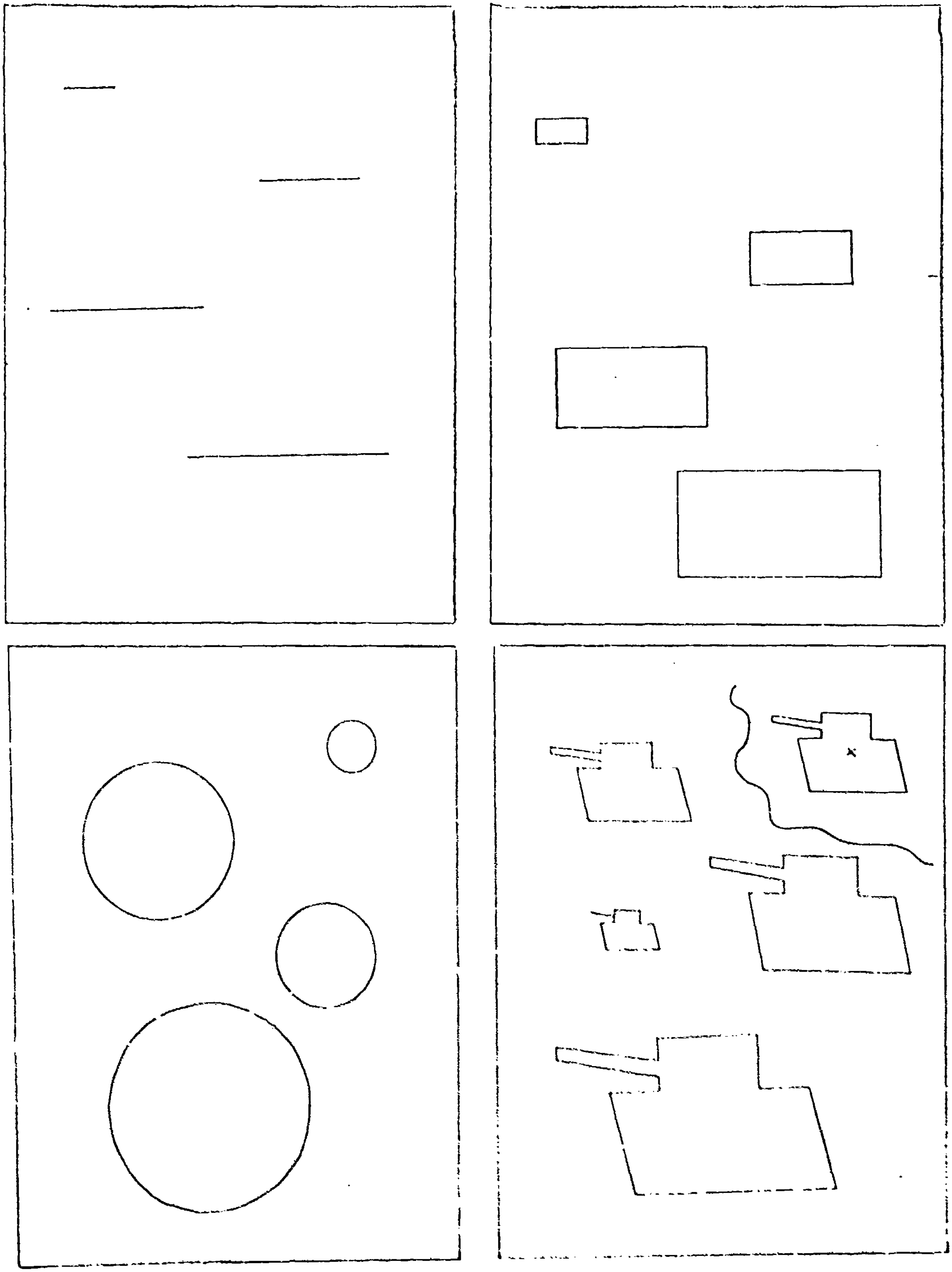


Figure 1

Target Pages (Reduced) for Each of Four Targets

(The aim point reference is shown on upper right of target page for tanks.

NOMINAL SIZE (in.)	LINE	RECTANGLE	CIRCLE	TANK
1	.0153* (1.53)†	.0252 (2.51)	.0267 (2.67)	.0400 (4.00)
2	.0325 (1.63)	.0316 (1.58)	.0241 (1.20)	.0672 (3.36)
3	.0333 (1.11)	.0500 (1.67)	.0622 (2.07)	.0975 (3.25)
4	.0425 (1.06)	.0616 (1.54)	.0772 (1.93)	.1809 (4.52)
AVERAGE ABSOLUTE ERROR	.0309	.0421	.0475	.0964

*mean absolute error in inches

†error as a percent of the nominal size

Table 1

Absolute and Percentage Errors for Four Targets

Appendix G

Handbook for Early Detection of Unit Learning in Operational Testing

Table of Contents

page

I. Introduction 76

II. Test Design Procedure 79

III. Testing for Learning 81

A. General Approach to Quantitative Analysis 81
B. Computation Procedure 83
C. Interpretation of Results 86

IV. Example 90

A. Basis Tracking Model 90
B. Project Stalk 93

I. Introduction

Operational tests on new weapon systems are conducted by Army units of the type that will eventually use the equipment. The purpose of this testing is to obtain results which are as realistic as possible, even though the systems themselves are still developmental. These test results are normally compared to results of similar tests with units using currently issued standard equipment. Since the comparison is made using two differently equipped troop units, an evaluation of the state of training of these two units must be made to ascertain that both units are at the same fully trained level. If this evaluation is not made, comparison of the two alternative systems may reflect training differences rather than improved system performance. The purpose of this manual is to assist the field commander to quantitatively evaluate the training level of his unit.

The performance curve describing the progress of learning is an asymptotic curve, often called a learning curve. Learning, described by such a curve, increases the unit performance at a decreasing rate. Performance level approaches a level, beyond which it probably will not go. This is the asymptote of the curve. When the crew or unit approaches this value the rate of learning approaches zero and we say the unit is fully learned and may conduct the tests. Such a performance curve is illustrated in Figure 1.

A complementary curve, called a learning curve is one in which the time to perform an activity is plotted against successive trails. Such a curve is illustrated in Figure 2. As Figure 2 indicates, this learning curve also approaches an asymptote. The asymptote may be zero or some minimum time.

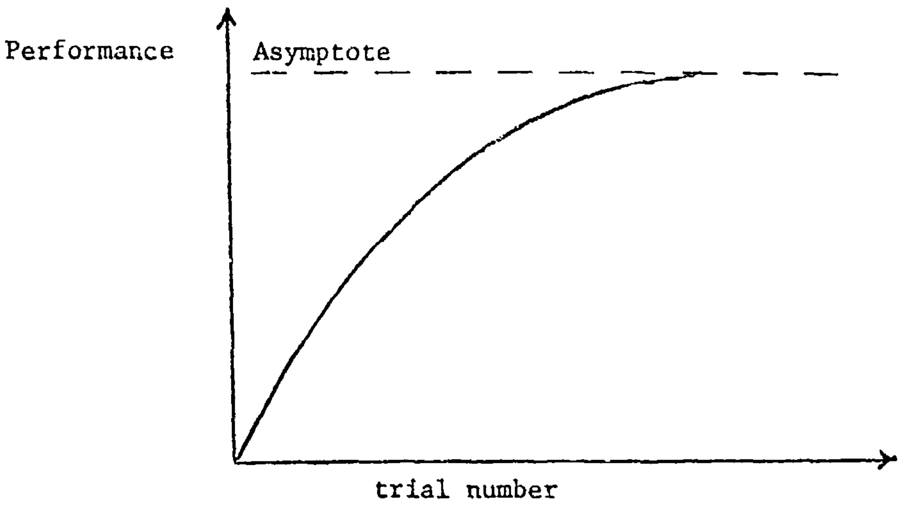


Figure 1. The Performance Curve

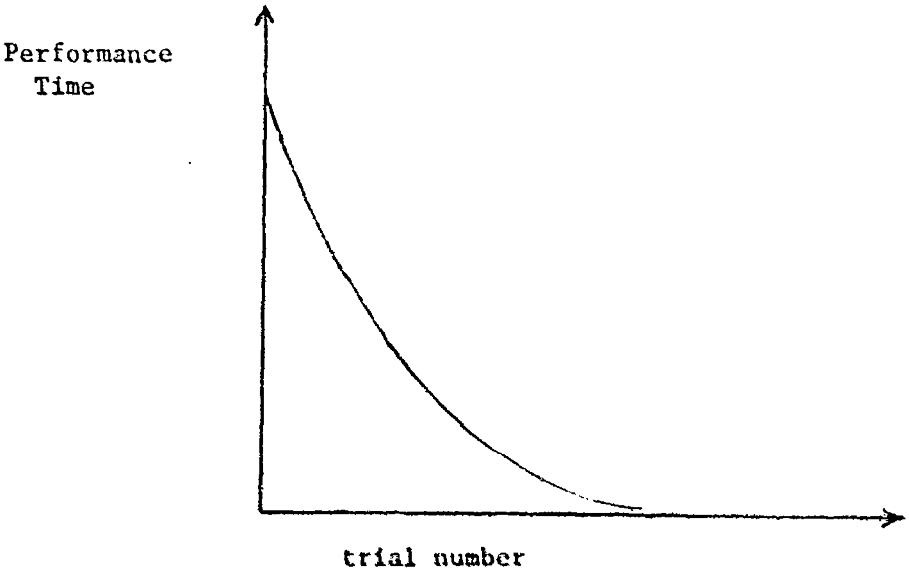


Figure 2. The Learning Curve

The theory of learning has been developed for individuals but little has been done to extend this theory to group learning. Based on limited research evidence, this manual provides a means for the unit commander to extend well known individual learning concepts to group learning.

An equation describing the performance curve relationship illustrated in Figure 1 is $Y = c - at^{-b}$, where Y is a measure of unit performance which increases as the unit becomes more skilled. Such a measure might be number of hits per trial. The parameter a in the learning curve model is determined by the cycle time at the beginning of the conditional learning phase. b is a parameter which is dependent on the rate of learning. c is the limit that is approached when complete learning has occurred. t is the time variable which may be either normal time or the trial number. This expression approaches an asymptote of c, which would represent a fully learned condition.

The corresponding equation describing the curve in Figure 2 is $Y = at^{-b}+c$, where Y now describes the performance in terms of a measure which decreases as the unit becomes more skilled. Such a performance measure might be time to acquire a target for example. The remaining symbols are the same as those for the performance curve of Figure 1.

While many other mathematical expressions describing learning may be found in the literature, current research indicates these two complementary models satisfactorily explain the type of learning involved in the training of army units. Therefore, no other models will be presented in this manual. The procedure which will be described is based on these models.

Learning can be divided into two phases, threshold learning and conditional learning. The first is that learning which occurs prior to the time an operation can be performed from memory. Conditional learning is learning which occurs after the unit can perform from memory, or without relying on a trial and error procedure. This manual deals with the second, or conditional learning phase.

Threshold learning is that learning that is accomplished through training films, lectures, demonstrations, etc. These involve the basic introduction to the hardware. The amount of such training needed depends on the complexity of the system. In general it involves individual learning rather than team or group learning. That is, it is the period during which each individual on the team learns his own part in the operation of the equipment. This is the reason it is not considered in this handbook.

II. Test Design Procedures

When learning is or may be a factor in the evaluation of operational test results, this should be taken into consideration when the test is designed. The tests must be designed in such a way that one or more appropriate performance measures are recorded in a time sequence. Time may be recorded as trial number or as clock time. Theoretically, if trial number is used, the models presented in the first section are not correct in that they are continuous models and trial number is a discrete variable. However, if we consider the increase in skill evidenced by the performance on successive trials as the increase in skill during the time interval between trials, the model will be approximately correct.

Tests should be designed in such a way therefore, that learning, if present, may be detected by the procedure discussed in this manual. Test trials should be close enough together in time that a loss of skill due to forgetting key points learned in training does not occur. If a unit has not recently used a weapon system, it should be retrained or retested on that equipment before the operational test may start. Recall that this model considers only the conditional learning phase, so a unit must be at that stage before evaluation may begin.

One thing that leads to the remission of skills by an army unit is personnel turnover. The army has been notorious in the rate of turnover occuring in its units. Thus if trials are separated by a lengthy time period the unit personnel should be checked to see if turnover has occured. If the unit has substantially different personnel during different trials the procedures discussed in this manual may be inappropriate. The unit training may need to be restarted whenever such turnover occurs.

The procedure may also be inappropriate if equipment changes between trials affect the unit performance. The procedure assumes that all personnel and equipment are the same, or at least any changes do not affect the performance, for each trial.

Often the performance will be measured against trial number. In some cases this may not be appropriate. Examples of such cases are those for which the trial length may differ among the trials. For example a crew may train for one hour in the morning and three hours in the afternoon or the next day. In this case training time may be more appropriate. It would also be more appropriate if training is conducted continuously.

Forgetting may create problems in some situations. This will be particularly important when there is a long time between trials, for example six months to a year. This forgetting may be caused by personnel or equipment turnover as previously discussed or, in the case of a relatively long time between trials, the loss of skills by the unit personnel. If this occurs the unit should be considered a new, untrained unit and the evaluation procedure started over.

If the personnel and equipment are the same and forgetting between trials is still present, performance may again be measured against time rather than trial number. This would allow the analyst to take this forgetting phenomonom into account in his evaluation.

Examples of the use of the algorithm for all of these situations may be found in the last section of this manual. The next section contains a step-by-step discussion of the procedure.

III. Testing for Learning

This section presents an overview of the approach used to investigate whether a change in performance is occurring during testing, followed by a detailed description of the computations and procedures required to quantitatively assess the data. A discussion of appropriate interpretations of the results of the quantitative procedure is also given.

A. General Approach to Quantitative Analysis

As discussed in earlier sections it is reasonable to assume that the model of the relations between performance and the amount of experience (measured by trials or time) of the unit is

$$Y_i = c - at_i^{-b}$$

where Y_i is the performance of the unit at the i^{th} trial.
a, b and c are parameters which depend on the nature of the task being performed. And t_i is time since the start of the learning or the number of trials since the start of learning.
A complementary expression is

$$Y_i = at_i^{-b}+c$$

where Y_i is the time to complete the task at time or trial t_i.

If the values of a and b were known, the rate of learning at any time, say t_k, could be found by computing the difference between Y_k and Y_{k-1} and dividing by $(t_k - t_{k-1})$. If this value were sufficiently small we could conclude that little learning was occurring. However, the values of a and b are not known for any potential situation and thus must be estimated. These estimates are based on the pairs of values (Y_i, t_i) observed in the testing.

Since learning is assumed to be non-linear across time, the relationship between the stated ratio and the parameter b is not monotonic. That is, for a particular value of t_k the ratio does not monotonically increase as b increases. When concern is with the rate of learning at a particular time, say t_k, the ratio is an appropriate measure. When concern is with the rate of learning over several trails this ratio must be examined for all trails; no one single value of this ratio can be considered. Note that since $Y_k - Y_{k-1} = a(t_k^{-b} - t_{k-1}^{-b})$, the ratio

$$\frac{Y_k - Y_{k-1}}{t_k - t_{k-1}}$$

may be written as

$$\frac{a(t_k^{-b} - t_{k-1}^{-b}}{t_k - t_{k-1}}$$

This indicates that if either a is small or b large, very little learning has occurred in the time interval t_{k-1} to t_k. Furthermore, there is no effect of c, the asymptote on the rate of change of learning.

An additional problem is that there is random variation in the outcomes of the tests. That is, even if all conditions, including the training level of the unit, were held as constant as possible, and the test repeated several times, the test results would not remain constant, but would vary randomly within some interval. Thus the result of the test at time t_k may not be exactly Y_k as computed from the above model, even if the true values of a, b and c were available. The greater this variability the more difficult it is to correctly identify the presence of significant learning.

Because of the above problems, the steps necessary to quantitatively detect the presence of significant changes in performance are

1. Estimate the rate of learning across trials
2. Estimate the variability of the observations
3. Establish a criterion so that the above estimate can be used to made a decision concerning learning.

B Computational Procedures

The computations required are easily carried out in conjunction with the worksheet shown in Figure 3. The following steps are necessary.

1. Record the values of the amount of experience at which the tests were conducted in column 1. This may be either the number of trials (that is, 1, 2, 3, etc.) or the number of hours or days of training.
2. Record the test result associated with each t_i in the corresponding row of column 4.
3. Add the entries in column 1, and divide this sum by the number of observations in the column, N.
4. Compute an entry for each row in column 2 by subtracting the result found in step 3 from each entry in column 1.
5. Compute entries for each row in column 3 by multiplying the corresponding entry in column 2 by itself.
6. Compute the entries in column 5 by subtracting the value in each row of Column 4 from the value in the next row of column 4. Note that there will be no entry in the last row of column 5.
7. Add the entries in column 5 and divide by the number of entries, (N-1).

FIGURE 3

COMPUTATIONAL WORKSHEET
FOR TESTING FOR LEARNING

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
i	t_i	$t_i-\bar{t}$	$(t_i-\bar{t})^2$	Y_i	X_i	$X_i-\bar{X}$	$(X_i-\bar{X})^2$	$(t_i-\bar{t})Y_i$
1								
2								
3								
4								
.								
.								
.								
.								
N								
	h		f				d	e

$\bar{t}$ = (Sum of entries in Column 1) divided by N = h/N

$X_i = Y_{i+1} - Y_i$

$\bar{x}$ = (Sum of entries in Column 5) divided by (N-1)

Compute

$$A = {}^{(e)}/_{(f)}$$

$$B = \frac{(N-1)(d)}{2N\ (N-2)}$$

$$C = \sqrt{\frac{12\ B}{N(N^2+1)}}$$

$$D = {}^{A}/_{C}$$

Plot D on Figure 4

8. Compute the entries in column 6 by subtracting the quantity found in Step 7 from each entry in column 5.

9. Multiply each entry in column 6 by itself to get the corresponding entry in column 7.

10. For each row, multiply the value in column 2 by the entry in column 4, and record the result in the corresponding row of column 8.

11. Add the entries in columns 3, 7 and 8.

The above steps complete the table. The values in the table are used to compute the following.

12. Divide the sum of the entries in column 8 by the sum of the entries in column 3. Call this result A.

13. Multiply the sum of the entries in column 7 by (N-1) and divide this result by the following product, 2 x N x (N-2). Call this result B.

14. Multiply the result in step 13 by 12 (not the result of Step 12) and divide by the following product N x $(N^2 + 1)$.

15. Take the square root of the quantity found in Step 14. Call this result C.

16. Divide the result found in Step 12 by the result found in Step 15. Call this result D.

The result found in Step 16 is the ratio of the estimate of the average rate of learning across trials to the variability of this estimate. By comparing this number to certain critical values a decision is made concerning learning. The critical values are based on statistical analyses and are dependent on two parameters. The first of these is the number of observations in the data. The more observations there are, the less likely it is that large values of the ratio computed will be observed when no learning is occurring. The second parameter is the level of protection we desire against deciding that learning is occurring when it is not. Allowing this risk to increase will cause us to conclude that learning is occurring for smaller values of the computed ratio. Appropriate values for this risk in most settings are in the range of 1% to 10%. The critical values are presented in Figure 4 and 5 for a risk of 5%. This figure is developed from tables of the Student t distribution. Figure 4 is to be used when testing for learning, Figure 5 when testing for increases in performance. If risk probabilities other than 5% are desired, the user should consult appropriate tables of the Student t distribution, available in most introductory Statistical tests. Such a table is included as Figure 6. This table is excerpted from Fundamental Concepts in the Design of Experiments, by Charles R. Hicks. If we are testing for learning the value of D should be less than the negative of the corresponding critical value in Figure 6 to conclude that learning is present. If we are testing for performance the value of D should exceed the value in Figure 6 to conclude that learning is present.

C. Interpretation of Results

There are two risks involved in arriving at decisions concerning the learning of the units. It might be concluded that the learning rate is significant when it is not. The probability of this occurring is the risk level associated with the critical value. Call this risk α. On the other hand, if the computed ratio is close to zero, there is no evidence that learning is occurring. There is a risk that this conclusion is wrong. In fact, when this occurrs the usual interpretation

Figure 4 Critical Values For Testing For Learning With an α risk of 5%

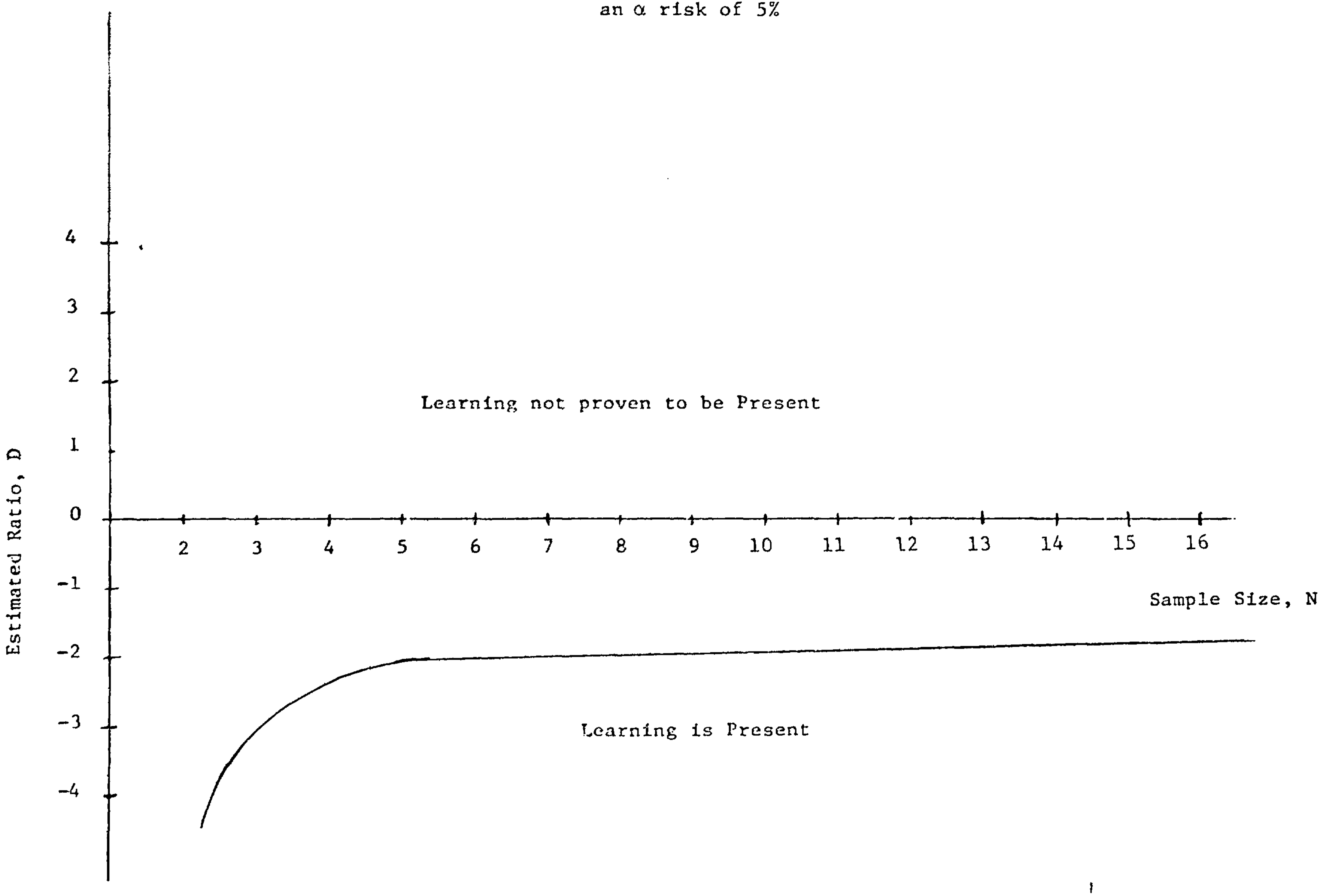


Figure 5 Critical Values For
Testing For Performance Improvement
With an α risk of 5%

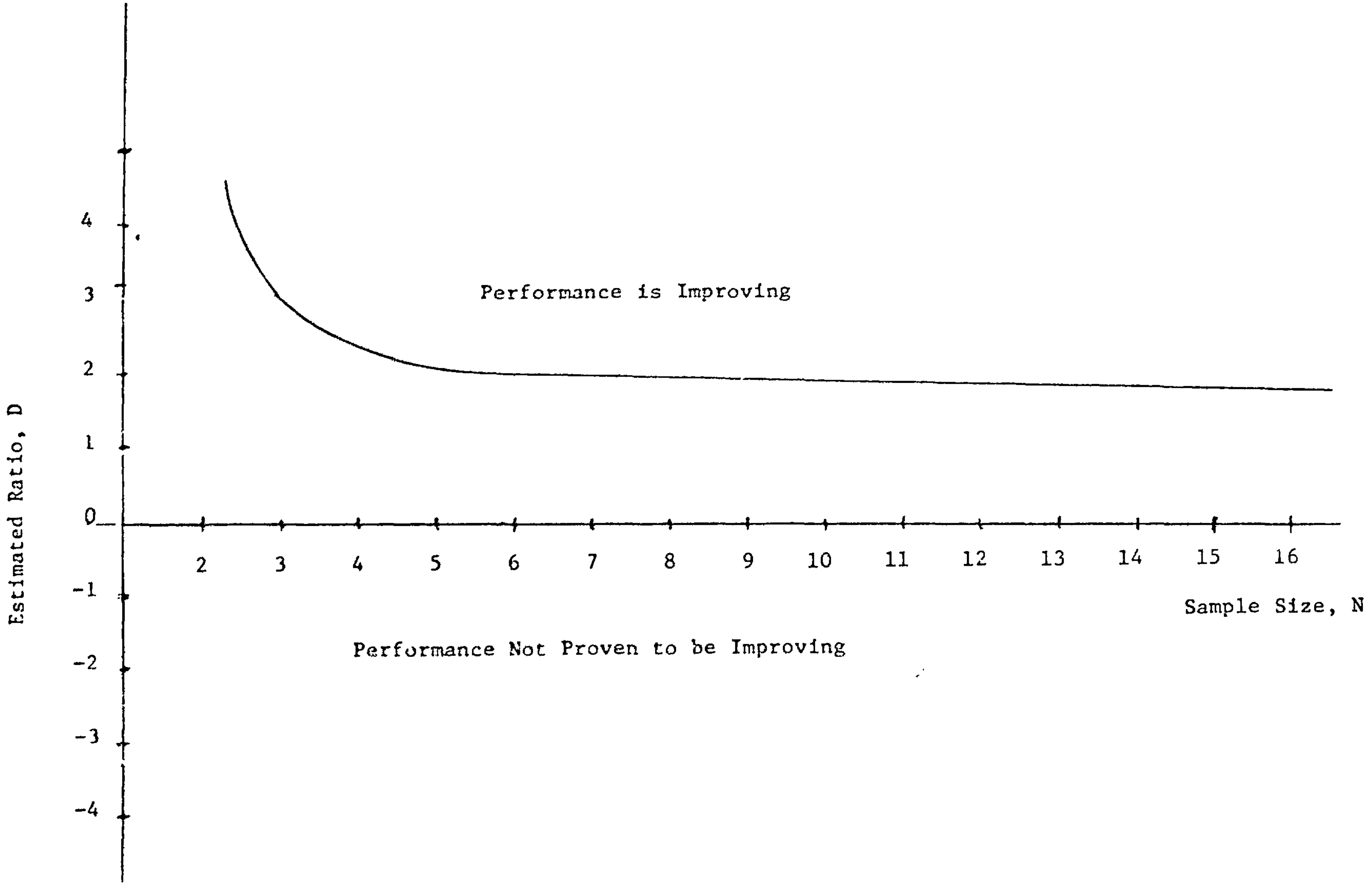


FIGURE 6. CRITICAL VALUES FOR TESTING

	RISK LEVEL			
N	1%	5%	10%	20%
1	31.82	6.31	3.08	1.38
2	6.96	2.92	1.89	1.06
3	4.54	2.35	1.64	.98
4	3.75	2.13	1.53	.94
5	3.36	2.01	1.48	.92
6	3.14	1.94	1.44	.91
7	3.00	1.90	1.42	.90
8	2.90	1.86	1.40	.89
9	2.82	1.83	1.38	.88
10	2.76	1.81	1.37	.88
11	2.72	1.80	1.36	.88
12	2.68	1.78	1.36	.87
13	2.65	1.77	1.35	.87
14	2.62	1.76	1.34	.87
15	2.60	1.75	1.34	.87

is that there is insufficient evidence to conclude that learning is occurring, but not that there is evidence that learning is not occurring. While this distinction is subtle, it is important. The risk associated with this conclusion, call it β, is related to two factors -- the number of observations and the risk level associated with concluding that learning is occurring, α. As the number of observations increases the β risk will decrease if α is held constant. As α is increased the β risk will decrease if the number of observations is held constant.

It is seen, therefore, that if the decision maker is to have confidence in his conclusion that learning is occurring, he would like to do it with as small an α risk as possible. On the other hand, if his conclusion is that learning is not significant, he is unsure as to exactly the risk he is taking, but is more confident of his result if the number of observations has been large or the α risk used was large.

The conclusion reached concerns the average rate of learning over all observations. It is not a measure of whether the unit is fully trained after the last trial, but rather is a measure of whether the individuals performed at the same level across all trials, which would be indicative of its being fully trained.

IV. Examples

The first example is presented in some detail to illustrate the computations required. Additional examples are then presented which illustrate the alternative model form and various interpretations of the results.

A. Basic Tracking Model

This example is based on actual experimental results in testing a viscous damped tripod. The performance measure was the standard deviation of the error (in the horizonal plane) from a marked aim point while tracking a moving target at constant velocity. The subject was familiar with tracking moving objects but had not previously operated this particular tripod.

The experiment was conducted by having the subject track the target as it was moved over a fixed course. This could be repeated as often as desired. Measurements on the actual aim point at various points along the course were made. The standard deviation of the distances from the marked aim point to the actual aim point was computed. Thus, there was only one performance measure for each time the target was run over the course. Since each of these runs should provide the same amount of experience to the subject, and since the length of time between repetitions was small it is appropriate to use the trial number as the measure of experience t_i.

Data were collected for the first six trials of a subject. These data are used to illustrate the procedure described in the previous section, following the steps suggested (See Figure 6).

Steps 1 and 2. The values of t_i and Y_i are entered in the table in columns 1 and 4. Note that the Y_i's are decreasing as experience is gained.

Steps 3 and 4. The sum of the entries in column 1, the t_i's, is 21. The average of these, 21/6 = 3.5, is substracted from each entry in column 1 to get the entry for column 2. For example, in the first row 1-3.5 = 2.5.

Step 5. Entries for column 3 are the entries in column 2 times themselves. For example, (-2.5) times (-2.5) = 6.25.

Step 6. The first entry in column 5 is found by subtracting the first entry in column 4 from the second entry in column 4. That is (3.49 - 4.21) = .72. The second entry in column 5 is (3.07 - 3.49) = -.42. The continues to row 5 where (2.23 - 1.81) = .42.

Steps 7, 8 and 9. We sum the entries in column 5 (being sure to keep track of signs) to get - 1.98 and divide by N-1 = 6 - 1 = 5 to get -0.396. Subtract -0.396 from -0.72 to get the first entry in column 6 of -0.324, multiply this by itself to get 0.105. Repeat this for each row to complete columns 6 and 7.

Step 10. Multiply the first entry in column 2 times the first entry in column 4 to get the first entry in column 8, that is, (-2.5) X 4.21 = -10.25.

Step 11. Add column 8, being sure to keep tract of signs. Divide the sum of column 8 by the sum of column 3.

FIGURE 7

COMPUTATION TABLE FOR EXAMPLE IV.1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
i	t_i	$t_i-\bar{t}$	$(t_i-\bar{t})^2$	Y_i	X_i	$X_i-\bar{X}$	$(X_i-\bar{X})^2$	$(t_i-\bar{t})Y_i$
1	1	-2.5	6.25	4.21	-.72	-0.324	0.105	-10.525
2	2	-1.5	2.25	3.49	-.42	-0.024	0.001	- 5.24
3	3	-0.5	.25	3.07	-.86	-0.464	0.215	- 1.54
4	4	0.5	.25	2.21	-.40	-0.004	0.000	1.11
5	5	1.5	2.25	1.81	.42	0.816	0.665	2.715
6	6	2.5	6.25	2.23				5.58
	21		18.50		-1.98		.986	-7.90

Carrying out Steps 12-16 we find D = -5.76. Plotting the point, D = -5.76, N = 6, on Figure 4 the point is clearly in the "Learning is Present" region. Thus we would conclude that the subject is not fully learned during the conduct of these six trials. Our risk of saying this when it is not true is less than 5%.

Note that this does not say that the subject is not fully learned after six trials, but it does indicate that it would be inappropriate to use data from all six of the trials to estimate the performance of fully learned subjects on this tripod.

B. Project Stalk Example

The data used in this example were extracted from one crew involved in Project Stalk. This involved tank crews firing at targets under different conditions of tank and fire control. The performance measure of interest is the time to achieve a target hit. In order to control for the effects of the different targets and conditions the times were aggregated across these factors. Thus, each time the crew completed the entire set of targets and conditions an observed value of the performance measure was available, and since these could reasonably be assumed to afford the same amount of experience, the t_i's were just the number of trials from the start of testing.

The data for five trials are presented in Figure 8(a) and an abbreviated computation table presented. The computed ratio of -1.93 would result in a decision that learning was not significant if an α risk of 5% were used. However, if an α risk of 10% were used this decision would be reversed. The decision maker would probably not be comfortable with the decision that learning was not significant, due to the small number of observations. Additionally, if the raw data are examined it

FIGURE 8. STALK DATA ALALYSIS

(a) First 5 responses, one crew

i	t_i	Y_i	X_i	$(X_i-\bar{X})^2$	$(t_i-\bar{t})Y_i$
1	1	172	130	$(99.5)^2$	-344
2	2	42	2	$(-28.5)^2$	- 42
3	3	40	2	$(-28.5)^2$	0
4	4	38	-12	$(-42.4)^2$	38
5	5	50			100
				13331.00	-248

$\sum(t_i-\bar{t})^2 = 10$

$A = {}^{-248}/_{10} = 24.8$

$B = {}^{4(13331)}/_{30} = 1777.5$

$C = 12.81$

$D = {}^{-24.8}/_{12.81} = -1.93$

(b) 2nd through 5th responses, one crew

i	t_i	Y_i	X_i	$(X_i-\bar{X})^2$	$(t_i-\bar{t})Y_i$
1	1	42	2	$(-2/3)^2$	-63
2	2	40	2	$(-2/3)^2$	-20
3	3	38	-12		+19
4	4	50		$(-9\ 1/3)^2$	75
				87.95	11

$A = {}^{11}/_{5} = 2.2$

$$B = \frac{3(87.95)}{16} = 16.5$$

$C = 1.7$

$D = {}^{2.2}/_{1.7} = 1.28$

seems obvious that the first trial is much different from the remaining four trials. That is, almost all of the learning seems to have occurred during the first trial. The decision maker might decide to consider only the second through the fifth trials. This value of the computed ratio 1.28 (see Figure 8(b)) would not be significant with an α risk as large as 14%.